Collana CERIDAP
diretta da
Diana Urania Galetta

Lorenzo Cotino Hueso and **Diana-Urania Galetta**

(editors)

THE EUROPEAN UNION ARTIFICIAL INTELLIGENCE ACT

A Systematic Commentary

Editoriale Scientifica

collana

CERIDAP SERIES

Director
Diana-Urania Galetta

In a context where the evolution of Public Administrations is perpetually shifting from legal, regulatory, and jurisprudential perspectives, thereby influencing every organisation and activity, Public Administration must inevitably incorporate extensive and diverse knowledge, both specific and interdisciplinary, pertaining to organisational structure and the execution of their competencies.

This is necessary to achieve an administrative action that is increasingly efficient and impartial, as well as to understand the needs of society and social issues and the expectations of citizens towards public entities in the administered community.

In this perspective, the CERIDAP series, which was established in conjunction with the homonymic Interdisciplinary Research Centre on Public Administration Law at the University of Milan and is closely linked to the CERIDAP Journal (https://ceridap.eu), aims to provide comprehensive analyses of subjects that are relevant to all three pillars of administration (organisation, activities, and judicial protection) and are conducted from a multidisciplinary perspective.

The CERIDAP series indeed positions itself as a place for in-depth study and research on issues related to the functioning of Public Administration, from the perspective of so-called good administration.

THE EUROPEAN UNION
ARTIFICIAL INTELLIGENCE ACT

A Systematic Commentary

Lorenzo Cotino Hueso and Diana-Urania Galetta

(editors)

Editoriale Scientifica
Napoli

The English version (and translation) has been funded by Project "Algoritmic law" (Prometeo/2021/009, 2021-24 Generalitat Valenciana).

GENERALITAT
VALENCIANA
Conselleria de Educación, Cultura,
Universidades y Empleo

Via San Biagio dei Librai, 39
80138 Napoli
ISBN 979-12-235-0281-5

Directors

Lorenzo Cotino Hueso
Professor of Constitutional Law at the University of Valencia. Valgrai

Diana-Urania Galetta
Professor of Administrative Law at the University of Milan (La Statale), Director of CERIDAP

Authors

Alessandro Mantelero
Jacques Ziller
Juan Gustavo Corvalán
María Victoria Carro
Lorenzo Cotino Hueso
Alfonso Ortega Giménez
Ángel Gómez de Ágreda
Jesús Jiménez López
Leire Escajedo
Miguel Ángel Presno Linera
Luis Miguel González de la Garza
Fernando Miró Llinares
Mario Santisteban Galarza
Inigo De Miguel Beriain
Gal-la Barrachina Navarro
Andrés Boix Palop
Rosa Cernada Badía
Vicente Álvarez García
Adrián Palma Ortigosa
Ignacio Alamillo Domingo
Eduard Chaveli Donet
Pere Simón Castellano
María Loza Corera
Francisca Ramón Fernández
Wilma Arellano Toledo
Antonio Merchán Murillo
Estrella David Gutiérrez
Guillermo Lazcoz Moratinos
Ana Aba Catoira
Marco Emilio Sánchez Acevedo
Idoia Salazar
Miguel Ángel Liébanas
José Antonio Castillo
Agustí Cerrillo i Martínez
Juan Carlos Hernández Peña
F. Javier Sempere
Aurelio López-Tarruella Martínez
Gabriele Vestri

GENERAL INDEX

Authors in order of appearance, with credits 27

General content of the work (by chapters and authors) 29

Presentation and introduction 33

ARTIFICIAL INTELLIGENCE ACT AND ITS GLOBAL CONTEXTUALISATION, FROM IBERO-AMERICA AND EUROPE

The Artificial Intelligence Act: the European legislator's response to the challenges of Artificial Intelligence

I. Introduction 41

II. The European perspective 41

III. The modulation of the so-called risk-based approach in first-generation legislation 47

IV. Conclusions 54

The Council of Europe Convention on Artificial Intelligence versus the EU Regulation: two very different legal instruments

I. The content of the draft Council of Europe Framework Convention 59

II. The reasons for a Council of Europe Treaty on Artificial Intelligence 62

III. The instrument of the Framework Convention versus the instrument of the Regulation 66

IV. The limits derived from the respective competences of the Council of Europe and the European Union 69

V. The need to ratify the Council of Europe Treaty versus the direct applicability of the EU regulation 74

VI. In conclusion 76

Artificial Intelligence regulation from outside the European Union: regulatory impulses from other parts of the world and a view from Iberoamerica

I. *Introduction* 77

II. *Getting out of the pot in time and other regulatory challenges* 79

1. Unity vs. fragmentation: “Horizontal” or “vertical” approach? 81
2. Second. Mandatory vs. voluntary and the strengthening of collaboration 84
3. Third. A risk-based approach 88

III. *Regulatory impulses from other parts of the world* 90

1. Australia 90
2. Canada 92

IV. *A Latin American perspective* 93

1. Mexico 93
2. Chile 95
3. Uruguay 96
4. Brazil 97
5. Colombia 99
6. Argentina 100

V. *Conclusion* 102

“ARTIFICIAL INTELLIGENCE”, TERRITORIAL SCOPE AND SCOPE OF THE REGULATION AND ITS RELATIONSHIP WITH DATA PROTECTION

What is “Artificial Intelligence” for the Regulation? Analysis, delimitation and practical applications

I. *The importance of the concept of Artificial Intelligence in the Regulation* 105

II. *The various definitions of Artificial Intelligence have been shuffled* 106

III. *The definition of Article 3 Regulation and its components: techniques, autonomy, adaptation, inputs and outputs, and context.* 109

IV. *Examples of systems that are or are not Artificial Intelligence* 113

The territorial scope of application of the Artificial Intelligence Act

I. *Approach* 115

II. *Application of Article 2.1 by economic operators* 118

III. Implementation of Article 2.1 by competent national authorities 121
1. The need for international cooperation 121
2. Harmonisation of regulatory standards 122
3. Interaction with International Law 123
IV. Application of Article 2.1. by the Courts of Justice 124
V. Extraterritorial application of the European Artificial Intelligence Act 125
VI. Final reflection 129

The exclusion of national security, defence, and military Artificial Intelligence systems from the Regulation and the applicable law

I. Introduction 131
II. Ethical codes for general-purpose Artificial Intelligence 133
III. Dual use of technologies 134
IV. The CCW Decalogue of Ethical Principles for Lethal Autonomous Weapon Systems 135
V. Significant differences between Civilian and Military uses of Artificial Intelligence 138
VI. Applicability of International Law to artificial intelligent systems 139
VII.Accountability and meaningful human control 140
VIII. Use of Artificial Intelligence in Security and Defence under the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law 142
IX. Conclusions 143

The Artificial Intelligence Act and the General Data Protection Regulation

I. Introduction 145
II. Artificial Intelligence systems and the processing of personal data 146
1. Artificial Intelligence Systems 146
2. AIS lifecycle and value chain 147
3. Processing of personal data and AIS 150
4. Identification of controllers of personal data in the AIS 153
III. The Relationship of the Artificial Intelligence Act and the General Data Protection Regulation 158
1. The need for a framework for the relationship between the two bodies of law 158
2. Application of the Regulation without prejudice to the application of the GDPR 161
3. The Regulation as "lex specialis" for policing purposes 162
4. Specific cases of legal basis for the processing of personal data in the Regulation 164

5. Independent personal data protection supervisory authorities 167
6. Human surveillance and automated individual decisions: Article 22 GDPR and the Regulation. 170
7. The right to explanation. Individual decisions in the context of certain high risk AIS and Article 22 GDPR 170
8. Collaboration of the Regulation in compliance with the GDPR 174
9. Limitations on collaboration by the initial regulatory space 176

IV. *Final reflections* 177

ARTIFICIAL INTELLIGENCE PROHIBITED OR UNACCEPTABLE FOR THE REGULATION (ARTICLE 5)

Biometric recognition in the Artificial Intelligence Act: Exemptions, prohibitions and high-risk specialties

I. *Biometric recognition in the Artificial Intelligence Act: key aspects of its regulation* 181

1. The regulatory approach to biometric recognition in the Regulation 181
2. Relevant background: the gap between the White Paper and EP Resolutions on some modalities of biometric recognition 184
3. Regulatory scheme of biometric recognition technologies in the Regulation 186

II. *Some key concepts in the typification of biometric recognition modalities in the statement of the Regulation: biometric data and biometric verification* 189

1. A new notion of “biometric data” to clearly encompass non-differentiating biometrics? 189
2. Common minimum of all biometric recognition systems 196
3. Biometric verification, outside the Regulation? 204
4. Non-remote identification, in limbo? 207
5. Is biometric screening covered? 209

III. *Biometric recognition systems affected by the prohibitions and restrictions of Article 5 of the Regulation: social assessment, prediction of criminal dangerousness, extension of facial recognition databases, inference of emotions in certain contexts and categorisations on specially protected data under Article 9 RGPD.* 209

1. Systematics of Article 5 regarding biometric recognition 209
2. Biometrics affected by the prohibitions of Art. 5.1 (C, D and E) of the Regulation: personality assessments for the purposes of citizen scoring, crime risk prediction and extension of facial recognition bases. 213

3. Biometric emotion recognition in workplaces and educational establishments, except for medical or security purposes (art. 5.1. f). Concept of emotion recognition in the Regulation 214
4. Biometric categorisation in order to deduce or infer their race, political opinions, trade union membership, religious or philosophical convictions, sexual life or sexual orientation (art.5.1 G). 216
IV. *Biometric recognition systems affected by the prohibitions and restrictions of Article 5 Regulation (and II): "real-time" remote biometric identification in public access areas for law enforcement purposes* 218
1. Global concern about remote biometric identification in public access spaces 218
2. Interpretation of the key concepts of the prohibition in Art. 5.1.h) 220
3. Paragraphs 2 to 8 224
V. *Biometric recognition modalities classified as high risk* 226
VI. *Large-scale recognition systems operational before the entry into force of the Regulation (Art. 111 and Annex X)* 228
1. Art. 111 and Annex X of Regulation 229
2. The Interoperability Regulations 230
3. Biometric recognition systems covered by Annex X: some relevant data 233
VII. *Final reflections* 237

The prohibition of Artificial Intelligence systems that evaluate and classify people based on data that are unrelated to the context in which they were generated and that lead to discrimination.

I. *Introduction* 241
II. *The Chinese social credit system* 244
III. *The development of qualification systems as a way of expanding surveillance capitalism* 247
IV. *The prohibition of certain systems that evaluate or classify natural persons* 251

The content of the so-called "subliminal techniques" and the vulnerabilities of specific groups of people in the Artificial Intelligence Act

I. *Introduction* 257

II. Developments in processing and content 258
III. What are subliminal techniques? 267
IV. An Artificial Intelligence that processes everything 274
V. Addictive technologies. AI action on groups, minors, young people, and other groups. 278
VI. Conclusions 282

The remaining Artificial Intelligence systems prohibited or unacceptable in the Artificial Intelligence Act

I. Introduction 283
II. Definitions and terminology: biometric categorisation technologies, emotions, biometric data and facial recognition. 285
III. Unacceptable practices: purpose and content of the prohibition 289
1. Easy recognition via scraping or non-selective extraction of facial images from the Internet and closed-circuit television (CCTV) 289
2. The prohibition of the use of Artificial Intelligence systems to infer emotions in the Regulation 290
3. The use of biometric categorisation Artificial Intelligence systems that individually classify natural persons on the basis of their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical convictions, sex life or sexual orientation. 292
IV. The coexistence of the regulation foreseen in the AI Act with data protection law: an overlapping and compatible regulation? 294

HIGH-RISK ARTIFICIAL INTELLIGENCE SYSTEMS: DELIMITATION AND ANALYSIS OF CERTAIN AREAS

Scope and Delimitation of High-Risk Systems in the Artificial Intelligence Act

I. High-Risk System Status is Essential for the Regulation 301
II. A Warning: The Regulation of Systems as High-Risk by the AIA does not imply their legal entitlement 303
III. Artificial Intelligence Systems in dangerous products of Annex I 304
1. Artificial Intelligence system as a safety component or product in products subject to a third party "Conformity Assessment" 304
2. The provider must reasonably know whether its Product can be in Annex I 307

IV. High-Risk Systems pursuing the purposes of Annex III. 308
1. Systems having a substantial influence on decision-making for the Purposes of Annex III and the criteria for considering it 308
2. Profiling with Artificial Intelligence for the purposes of Annex III shall always be High-Risk 313
3. The purposes of High-Risk systems in Annex III 314
4. Presumption that the Artificial Intelligence System pursuing Annex III purposes is indeed High Risk. Special obligations and actions 321
5. When the deployer alters a system to a High-Risk Purpose 322

V. The role of the Commission, Criteria, Delegated Acts, updating and amendment of High-Risk schemes, in particular from Annex III. 323

VI. Summary and Conclusions 325

The regulation of predictive
policing systems in the Artificial Intelligence Act

I. Introduction 327

II. Predictive Policing Systems and Artificial Intelligence 329
1. Police organisation in times of digitalisation and the “misnamed” predictive policing 329
2. The ethical risks of predictive policing (and those added by the use of Artificial Intelligence) 333

III. Regulating predictive policing in the Artificial Intelligence Act 340
1. Developments in the regulation of the use of predictive policing systems in the legislative process of the Artificial Intelligence Act 340
2. Predictive policing systems prohibited in the AI Act 343
3. “High-risk” predictive policing and its implications 346
4. Conclusions 348

The applicability of the Artificial Intelligence Act
to the health sector and specialities regarding its compliance

I. Introduction 351

II. Regulation of medical devices in the AI Act and their consideration as high-risk by Annex I or III 352
1. Preliminary analysis: the regulation of medical devices incorporating Artificial Intelligence in the AI Act 352
2. Systems which are high risk in accordance with the provisions of Annex III 352

3. Systems which may or may not be high risk as set defined in Annex I. 354
III. The regulation of medical devices: the MDR and IVDR provisions 355
1. Medical devices. A characterisation 355
2. Classes of medical devices and supervision requirements inherent to each type according to the MDR 358
3. IVDR and in vitro diagnostic devices 360
4. Exceptions to the general regime for the systems included in Annex III 361
IV. Recapitulation 363

The applicability of the Artificial Intelligence Act to the field of public administration and public services and special features regarding compliance: special attention to Annex III and administrative action and particularities of compliance

I. Introduction: the approach of the AI Act and the projection of its controls and guarantees on the actions of the public authorities 365
II. Development, processing and final content of the precepts of the Act that most affect the public authorities 372
III. Some conclusions on the application of the Act to the public sector 383

Large Artificial Intelligence platforms and systems for political influence: the intersection between the "Digital Services Act" and the Artificial Intelligence Act from a risk perspective

I. Introduction. Basis for the specific treatment of digital platforms and political influence systems in the Act. 387
II. A brief look at the "iter legis" of the Act on the regulation of large platforms and political influence systems 390
1. Specific attention in the Act to Artificial Intelligence systems for political influence 392
2. Algorithmic recommendation systems: treatment in the Artificial Intelligence Parliament proposal version and its rationale 394
III. The logic of risk on large platforms: the complementarity between the DSA and the AI Act 395
IV. Special features of the application of the Act to large-scale AI platforms and systems for political influence 400
1. Risk management safeguards 400
2. Algorithmic transparency safeguards: explainability vs. opacity 402
3. Procedural safeguards 406

4. Organic Guarantees and Digital Governance 408
V. *Conclusions* 410

GENERAL REGIME APPLICABLE TO HIGH-RISK ARTIFICIAL INTELLIGENCE SYSTEMS

The implementation of harmonised standards and common specifications in the field of Artificial Intelligence (Articles 40 and 41 AIA)

I. *Introduction* 415
1. The regulation of Artificial Intelligence through the harmonising technique of the new approach 415
2. A brief introduction to the basic elements of the new approach harmonisation technique as applied to Artificial Intelligence 417
II. *Harmonised standards* 419
1. A basic preliminary question: harmonised rules have the legal nature of Community law. 419
2. The distinction between European standards and harmonised European standards 420
3. The involvement of the European standardisation bodies and the Commission in the procedure for drawing up harmonised standards 423
4. Provisions on harmonised standards during the process of the Commission, the Council and the European Parliament during the procedure for the Proposal for a Regulation 426
5. Harmonised standards in the final text of the AI Act 430
6. The problems of applying standardisation techniques to the Artificial Intelligence Act in the EU 431
7. Current results of standardisation work in Artificial Intelligence 435
III. *The common specifications* 437
1. Some initial thoughts on its concept 437
2. The evolution of the regulation of the common specifications during the procedure of the Proposal for a regulation: from the Commission's draft to the European Parliament's amendments and the Council's compromise text. 439
3. The essential elements that make up the common specifications in the AI Act 442
4. Other technical solutions equivalent to those offered by harmonised standards and by the common specifications 446

Conformity assessment in the design and production of Artificial Intelligence-based systems in the context of the " New Legislative Framework".

I. Introduction 449

II. Evolution, processing, and final content of the articles of the Artificial Intelligence Act involved 450

III. Conformity assessment in European Union legislation 451

IV. The forms of conformity assessment in the Artificial Intelligence Act 453

1. Conformity assessment procedure based on internal control 453
2. The assessment carried out by a notified body 454

V. Conformity assessment according to the type of Artificial Intelligence system 457

1. Artificial Intelligence Systems whose purpose is considered high-risk (high-risk purposes) 457
2. Artificial Intelligence systems for products or safety components of products considered to be of high risk 459

VI. Notified bodies responsible for carrying out the conformity assessment of the different Artificial Intelligence systems. 463

VII. Cases where conformity assessment is not required or presumptions of conformity of compliance exist. 465

1. Prior authorisation to place Artificial Intelligence systems on the market 465
2. Placing the Artificial Intelligence system on the market without prior authorisation 466
3. Exemptions from conformity assessment for products subject to harmonisation legislation 467
4. Presumptions of Conformity 467

VIII. Reflections on the conformity assessment regulation set out in the AI Act 468

IX. Conclusions 469

General regime of obligations for providers and deployers in the Artificial Intelligence Act

I. Introduction 471

II. Development, processing, and final content of the articles of the Act involved 472

III. Operators in the value chain of Artificial Intelligence systems in the Act 472

1. The provider and his obligations 473
2. The deployer and his obligations 474

3. The authorised representative and his obligations 481
4. The importer and his obligations 482
5. The distributor and his obligations 483
6. Possible alterations to operators' responsibilities 485
IV. *The notifying authorities* 486
1. Concept of notifying authority 486
2. Activities of the notifying authority 488
V. *Measures targeting small-scale providers and users* 490
VI. *Notification of serious incidents* 492
1. Concept of serious incident 492
2. Who, before whom, and when must it be notified? 493
3. Proceedings following the notification of the incident 495
VII. *Conclusions* 496

Subjects and actors in conformity assessments (notified bodies)

I. *Introduction* 499
II. *The notification procedure* 501
1. The application for designation 502
2. The notification procedure 504
3. Identification and publicity of notified bodies 505
4. Changes to the notification 506
5. Questioning the competence of notified bodies 508
III. *The performance of notified bodies* 508
1. Requirements applicable to notified bodies 508
2. Operational obligations of notified bodies. Coordination by the Commission 516
3. Issue and validity of certificates 517
4. Information obligations of notified bodies 519
IV. *Recapitulation* 520

THE OBLIGATIONS OF SUPPLIERS AND DEPLOYERS OF HIGH-RISK SYSTEMS

The fundamental rights impact assessment by deployers of Artificial Intelligence systems in the Regulation

I. *Introduction* 523
II. *Impact assessments as a tool for weighing up fundamental rights* 524
1. Differences between risk analysis and impact assessments. The example of data protection 526

2. Differences between the risks to be addressed in the risk management system of Article 9 vs. the fundamental rights impact assessment of Article 27 AI Act 528
3. Typologies of impact assessments 531
III. Background on fundamental rights assessments of AI systems prior to the AI Act 534
IV. The Fundamental Rights Impact Assessment of High-Risk Artificial Intelligence Systems in the Act 538
1. Development, processing and final content of the articles of the Act involved 538
2. Analysis of FRIA in the Act 542
V. Steps to be taken in a FRIA 550
VI. Fundamental rights impact assessments and data protection impact assessments 561
VII. Recapitulation and conclusions 564

Risk management systems as a specific obligation for high-risk Artificial Intelligence systems in Article 9 of the Regulation

I. What is a risk management system? Conceptual autonomy with respect to related figures 567
1. Risk identification (or assessment phase) 568
2. Rights risk assessment and its differentiation from provider risk analysis and risk management 569
3. Risk mitigation, monitoring and review 577
II. Evolution of the meaning, content, and recipients of the obligation to have a risk management system (Article 9 AIA) 580
1. What is a risk management system according to the Regulation? The content of the obligation 584
2. Obligated subjects. Who is obliged to have a risk management system? Summary table of obligations related to high-risk systems 588
III. Recapitulation and conclusions 592

Data and data governance and connections to data protection principles in Article 10 of the Artificial Intelligence Act

I. Introduction 595
II. Data governance 596
1. Concept of data governance 596
2. European context 599
3. Concept of governance in the field of Artificial Intelligence 601

III. Development, processing and final content of Article 10 603
1. Roles involved 603
2. Obligations 605
IV. Critical approach 610
V. Confluence of data protection regulation 621
VI. Conclusions 625

Quality management systems, technical documentation and documentation keeping in the Regulation

I. Introduction 629
II. Article 17 of the AIA on the quality management system 634
III. Article 11 of the AIA on technical documentation with Annexes 639
IV. Article 18 of the AIA on documentation keeping 644
V. The initial Article 50 of the AIA on document retention 644
VI. Conclusions 645

The obligation to keep records of high-risk systems in the Artificial Intelligence Act

I. Introduction 649
II. Some preliminary notions related to record-keeping obligations 650
III. Registration obligations in the Regulation: evolution, processing and final content 655
IV. Final reflections and analysis 662

Transparency and provision of information to deployers in Article 13 of the Artificial Intelligence Act

I. Introduction: a general approach to Article 13 of the Regulation 665
1. The legislative process of shaping Article 13 and its interpretative challenges 666
2. Objectives and methodological considerations 668
II. Systematic description of Article 13: transparency dimensions 673
III. Transparency, interpretability and explainability: their (a)systematic treatment in Article 13 679
1. Meaning and types of transparency in the regulation and in Article 13 682
2. Interpretability in Article 13: discouraging black box models? 685
3. Explainability in Article 13: an ambiguous and limited approach 690
IV. Subjective and formal scope of Article 13 698
1. Subjects and the purposes of transparency in Article 13: the regulation's major absences 698

2. The formal scope of transparency: the undefined "appropriate type and level of transparency". 700
V. The material content of the transparency obligation 704
1. Information on functional suitability and other properties 706
2. Information on functional correctness or predictive performance: "accuracy" and its "metrics". 708
VI. Final assessment of Article 13 of the Regulation 713

Human oversight or monitoring in Article 14 of the Artificial Intelligence Act: a mere mandatory requirement for high-risk systems?

I. Introduction 717
II. Human oversight or monitoring in the Commission's April 2021 proposal 719
1. Human supervision to prevent or reduce risks 720
2. Effective human oversight by design 721
3. How to achieve an effective human supervision? 722
4. The role of the deployer in human oversight 723
III. The trajectory of human oversight in the ordinary legislative procedure 725
1. Voices calling for the right to human intervention (and other safeguards) for decision-making based on high-risk systems 726
2. Humans, what humans? 728
IV. The final version of human oversight in the AIA at a glance 730
V. Some reflections on what we can expect from human oversight in the AI Act 730
1. Can human beings fulfil the normative purpose of human oversight in the Regulation? 731
2. Is it necessary, under the Regulation, to have humans in the loop of high-risk AI decision-making to ensure the required effective human oversight? 733
3. Beyond human supervision, do we have a human-centred Regulation? 735
VI. Conclusions 737

Accuracy and robustness of high-risk Artificial Intelligence systems in Article 15 of the Artificial Intelligence Act

I. Introduction to accuracy and robustness in high-risk AI systems 739
II. Development, processing and final content of Article 15 745
III. The requirements of an adequate level of accuracy and robustness 749
1. Metrics and system performance 751

2. Assessment of accuracy and robustness to ensure the quality of the system 758
IV. Conclusions 759

Cybersecurity in high-risk Artificial Intelligence systems in Article 15 of the Artificial Intelligence Act

I. Does cybersecurity apply to all Artificial Intelligence systems? The Cybersecurity Obligation from the Proposal to the Final Approval 732
1. Evolution, processing and final content of cybersecurity as a requirement in AI systems in the framework of the adopted proposal 764
2. Cybersecurity in high-risk AI systems 771
II. The European Cybersecurity Certification Framework as an assurance tool for high-risk AI systems 775
1. Cybersecurity certifications in high-risk AI systems, mandatory or voluntary? 776
2. Cybersecurity obligations in high-risk AI systems 778
III. Conclusions 783

Post-market monitoring on high-risk AI systems in the Artificial Intelligence Act. Description, measures and use cases

I. Post-market monitoring in the Regulation 785
1. Introduction 785
2. What is a post-market monitoring plan and what does it include? 786
3. Why is a post-market monitoring plan necessary? 789
4. Post-market monitoring plan in the Regulation 790
5. Who should conduct the post-market monitoring system? 791
II. Addressing the Post-Market Monitoring requirement 792
1. Monitoring Plan and Monitoring System. Key elements of monitoring 792
2. Design of the Monitoring System 793
3. The Indicator concept 794
4. Measures to be developed in the Monitoring Plan 795
5. Validity of the Monitoring System and Plan 796
III. Conclusions 796

GENERAL-PURPOSE ARTIFICIAL INTELLIGENCE, NON-HIGH-RISK SYSTEMS AND ARTICLE 50 SYSTEMS

General-purpose Artificial Intelligence, foundational models (and "GPT Chat") in the Artificial Intelligence Act

I. Introduction 801
II. General Artificial Intelligence, general-purpose Artificial Intelligence, foundational models and generative Artificial Intelligence 803
III. What is and what is not general-purpose Artificial Intelligence in the Regulation? general-purpose models with systemic risk and exclusion of models specifically intended for research purposes 805
IV. Some regulatory challenges of generative Artificial Intelligence 809
V. Applicability of the Regulation as a general rule and regulatory developments in the treatment of general-purpose Artificial Intelligence 814
VI. Obligations of general-purpose Artificial Intelligence providers in the Regulation 816
VII.Supervision and monitoring, and sanctioning regime 820
VIII. Conclusions 822

Codes of conduct, seals or certifications for Artificial Intelligence systems that are not high risk (Article 95 of the AI Act)

I. The regulation in Article 95 of Codes of conduct for non-high-risk systems 825
II. Finally, the Regulation has not included binding principles for all types of Artificial Intelligence systems. 827
III. The insertion of Artificial Intelligence in certification models and technology seals in the EU: a Spanish Artificial Intelligence seal? 830
IV. The various European and international Artificial Intelligence initiatives and certification tools or labels 832
V. In conclusion 838

Article 50 of the AI Act and the transparency obligations for providers and deployers of certain Artificial Intelligence systems

I. Limited-risk Artificial Intelligence systems 841
II. Development, processing and final content of Article 50 the AIA 842
III. The scope of the transparency obligations provided for in Article 50 AIA 844
IV. Artificial Intelligence systems interacting directly with natural persons 845
1. Covered systems 845

2. Scope of obligations 846

V. *Artificial Intelligence systems that generate synthetic content* 847

1. Covered systems 847

2. Scope of obligations 849

VI. *Artificial Intelligence systems for emotion recognition* 850

1. Covered systems 850

2. Scope of obligations 854

VII. *Artificial Intelligence systems used for biometric categorisation* 855

1. Covered systems 855

2. Scope of obligations 857

VIII. *Artificial Intelligence systems that generate or manipulate content that constitutes deep fake* 858

1. Covered systems 858

2. Scope of obligations 861

IX. *Recapitulation* 862

SANDBOX, GOVERNANCE, OVERSIGHT, SANCTIONS, RIGHTS AND CONFIDENTIALITY IN THE REGULATION

Sandbox, controlled spaces and real-world testing of Artificial Intelligence systems in the Regulation. Measures for SMEs, startups and micro-enterprises

I. *The "Measures in support of innovation" of Chapter VI* 867

II. *Origin and concept of sandboxes and controlled spaces* 868

III. *Artificial Intelligence sandbox experiences* 870

IV. *The advantages of Artificial Intelligence sandboxes from different points of view* 874

V. *The Regulatory Framework of an Artificial Intelligence Sandbox under the Regulation* 878

VI. *Exceptionality of the legal regime and liability of participants. Theory, reality and Regulation* 881

VII. *Sandbox authorities, selection of participants, duration, development, and obtaining conformity assessment* 886

1. The competent sandbox authority and its cooperation with other national and European authorities 886

2. Selection and admission of participants and duration of the sandbox 887

3. Developing, completing and achieving a sandbox compliance assessment 889

VIII. *Data protection in the context of an Artificial Intelligence sandbox* 889

IX. Testing under high-risk Artificial Intelligence system conditions 892
X. SMEs, start-ups and micro-enterprises in the Regulation 894
XI. Conclusions and summing up 895

GOVERNANCE AND OVERSIGHT OF THE ARTIFICIAL INTELLIGENCE ACT: MARKET SURVEILLANCE AUTHORITIES, THE COMMISSION AND THE VARIOUS ENTITIES

I. Introduction 899
II. Development, processing and final content 900
III. The EU Governance Model of the AI Act 901
IV. European Commission. Powers and functions 903
V. AI Office. Nature, structure and functions 907
VI. Competences of Member States and competent national authorities 911
1. Market surveillance authorities 914
2. Notifying authority 915
VII. European Committee on Artificial Intelligence and the permanent sub-groups 916
1. European Committee on Artificial Intelligence. Structure and terms of reference 916
2. Standing sub-groups for market surveillance and notifying authorities 919
VIII. Other advisory, support and collaboration bodies 921
1. Advisory Forum 921
2. Scientific panel of independent experts 922
3. European Centre for Algorithmic Transparency and Artificial Intelligence Testing Support Structures 923
IX. Safeguard, market surveillance and control procedures for Artificial Intelligence systems in the Union 924
1. Procedure for Artificial Intelligence systems posing a risk at national level 925
2. Union safeguard procedure 927
3. Procedure in respect of compliant Artificial Intelligence systems presenting a risk 928
4. Procedure in case of formal non-compliance 929
5. Procedure for dealing with AI systems classified by the provider as non-high-risk in application of Annex III 929

THE SANCTIONING REGIME IN THE ARTIFICIAL INTELLIGENCE ACT

I. Introduction 931
II. Article 71 of the Regulation: the need for legislative development

and interaction with other legislation and alternatives to fines and specifications for administrations. 932
III. Prescription and classification of infringements and amount of fines in Article 71 935
IV. Criteria for determining the amount of the fine under Article 71 940
V. Proposals and changes by Parliament and in the final versions 945
1. Proposed changes in Parliament 945
2. Developments in the final texts 946
VI. Article 72 on the sanctioning powers of the European Data Protection Supervisor 949
VII. Conclusions. 951

RIGHT TO LODGE A COMPLAINT AND RIGHT TO AN EXPLANATION.
MEANS OF REDRESS FOR INDIVIDUALS IN THE ARTIFICIAL INTELLIGENCE ACT

I. Introduction 955
II. The right to lodge a complaint with a market surveillance authority 956
1. Developments in the text of the provision in the preparatory work 957
2. Differences between the regulation of the right to lodge a complaint in the AI Act and that established in the GDPR and other European digital laws 960
3. Regulation of the right to lodge a complaint in the Act 964
III. The right to an explanation of individual decision-making 967
1. Developments in the text of the provision in the preparatory work 968
2. The Explainability Principle for Artificial Intelligence Systems 969
3. Conditions for the exercise of the right to an explanation 972
4. The relationship between the right to an explanation of Article 86 and the GDPR 976
5. Limits to the right to an explanation: Intellectual property rights and confidential information 978
IV. The Role of Complainants and Collective Interest Associations 980
V. Conclusions 983

ACCESS TO DOCUMENTS AND CONFIDENTIALITY IN THE ARTIFICIAL INTELLIGENCE ACT

I. Introduction 985
II. Analysis of the content of Article 77 of the AIA 988
III. Analysis of the content of Article 78 of the AIA 991
IV. Conclusions 994

AUTHORS IN ORDER OF APPEARANCE, WITH CREDITS

Alessandro Mantelero, *Senior Lecturer in Civil Law at the Politecnico di Torino and holder of the Jean Monnet Chair in Mediterranean Digital Societies and Law.*

Jacques Ziller, *Professor of Public Law and European Union law, Universities Paris-1 Panthéon Sorbonne and Pavia*

Juan Gustavo Corvalán, *Director of the Innovation and Artificial Intelligence Laboratory of the Faculty of Law of the University of Buenos Aires.*

María Victoria Carro, *PhD Candidate, University of Genoa. Research Director, UBA IALAB*

Lorenzo Cotino Hueso, *Professor of Constitutional Law at the University of Valencia. Valgrai*

Alfonso Ortega Giménez, *Senior Lecturer in International Private Law at the Miguel Hernández University of Elche (Alicante).*

Ángel Gómez de Ágreda, *Lecturer at the Universidad Politécnica de Madrid. Spanish Ministry of Defence. Odiseia*

Jesús Jiménez López, *Director of the Council for Transparency and Data Protection of Andalucia*

Leire Escajedo, *Senior Lecturer in Constitutional Law University of the Basque Country/EHU*

Miguel Ángel Presno Linera, *Professor of Constitutional Law at the University of Oviedo*

Luis Miguel González de la Garza, *Reader in Constitutional Law UNED*

Fernando Miró Llinares, *Professor of Criminal Law and Director of the CRIMINA Centre, Miguel Hernández University, Elche.*

Mario Santisteban Galarza, *University of the Basque Country*

Iñigo De Miguel Beriain, *Ikerbasque research professor. Researcher at the University of the Basque Country/Euskal Herriko Unibertsitatea. Member of the Spanish Bioethics Committee*

Gal-la Barrachina Navarro, *University of Valencia*

Andrés Boix Palop, *Senior Lecturer in Administrative Law at the Universitat de València*

Rosa Cernada Badía, *Lecturer in Administrative Law. Catholic University of Valencia San Vicente Mártir*

Vicente Álvarez García, *Professor of Administrative Law at the University of Extremadura*

Adrián Palma Ortigosa, *Lecturer in the Department of Administrative Law of the Universitat de València*

Ignacio Alamillo Domingo, *PhD in Law*

Eduard Chaveli Donet, *Digital Law Specialist Attorney. Head of Consulting Strategy at Govertis, Part of Telefónica Tech.*

Pere Simón Castellano, *Senior Lecturer in Constitutional Law. International University of La Rioja – UNIR*

María Loza Corera, *PhD in Law. Lead Advisor at Govertis part of Telefónica Tech. Lecturer at the International University of La Rioja*

Francisca Ramón Fernández, *Professor of Civil Law at the Universitat Politècnica de València*

Wilma Arellano Toledo, *PhD from the Complutense University of Madrid. OdiseIA*

Antonio Merchán Murillo, *PhD. Lawyer. OdiseIA. Lecturer (habilitation as "Professor") at the University of Cadiz.*

Estrella David Gutiérrez, *Lecturer in Constitutional Law at the Universidad Complutense de Madrid*

Guillermo Lazcoz Moratinos, *Centre for Biomedical Research Network (CIBERER – IS-CIII)*

Jiménez Díaz Foundation Health Research Institute (IIS-FJD)

Ana Aba Catoira, *Senior Lecturer in Constitutional Law at the University of A Coruña*

Marco Emilio Sánchez Acevedo, *Lawyer. Lecurer and researcher at the Catholic University of Colombia.*

Idoia Salazar, *PhD. Lecturer at CEU San Pablo University. President of Odiseia*

Miguel Ángel Liébanas, *Criminologist expert in Intelligent Systems. Odiseia. CEO of Human Trends*

José Antonio Castillo, *PhD. Ramón y Cajal Researcher – University of Granada*

Agustí Cerrillo i Martínez, *Professor of Administrative Law at the Universitat Oberta de Catalunya (Open University of Catalonia).*

Juan Carlos Hernández Peña, *Senior Lecturer in Administrative Law at the University of Navarra*

F. Javier Sempere, *Supervision and Data Protection Director of the General Council of the Judiciary. PhD candidate at CEU International Doctoral School (CEINDO).*

Aurelio López-Tarruella Martínez, *Senior Lecturer in International Private Law at the University of Alicante*

Gabriele Vestri, *PhD in Law, Founder and President of the Public Sector and Artificial Intelligence Observatory.*

GENERAL CONTENT (BY CHAPTERS AND AUTHORS)

Presentation and introduction, by Lorenzo Cotino Hueso and Diana-Urania Galetta

Artificial Intelligence Act and its global contextualisation, from Ibero-America and Europe

The Artificial Intelligence Act: the European legislator's response to the challenges of Artificial Intelligence, *by Alessandro Mantelero*

The Council of Europe Convention on Artificial Intelligence versus the EU Regulation: two very different legal instruments, *by Jacques Ziller*

Artificial Intelligence Regulation from outside the European Union: regulatory impulses from other parts of the world and a view from Ibero-America, *by Juan Gustavo Corvalán and María Victoria Carro*

"Artificial Intelligence", territorial scope and scope of the Regulation and its relationship with data protection

What is "Artificial Intelligence" for the Regulation? Analysis, delimitation and practical applications, *by Lorenzo Cotino Hueso*

The territorial scope of application of the Artificial Intelligence Act, *by Alfonso Ortega Giménez*

The exclusion of national security, defence and military Artificial Intelligence systems from the Regulation and the applicable law, *by Angel Gomez de Agreda*

The Artificial Intelligence Act and the General Data Protection Regulation, *by Jesús Jiménez López*

Artificial Intelligence prohibited or unacceptable for the Regulation (Article 5)

Biometric recognition in the Artificial Intelligence Act: exemptions, prohibitions and high-risk specialties, *by Leire Escajedo*

The prohibition of Artificial Intelligence systems that evaluate and classify people based on data that are unrelated to the context in which they were generated and that lead to discrimination, *by Miguel Ángel Presno Linera*

The content of the so-called "subliminal techniques" and the vulnerabilities of specific groups of people in the Artificial Intelligence Act, *by Luis Miguel González de la Garza*

The remaining Artificial Intelligence systems prohibited or unacceptable in the Artificial Intelligence Act, *by Pere Simón Castellano*

High-risk Artificial Intelligence systems: delimitation and analysis of certain areas

Scope and delimitation of high-risk systems in the Artificial Intelligence Act, *by Lorenzo Cotino Hueso*

The regulation of predictive policing systems in the Artificial Intelligence Act, *by Fernando Miró Llinares and Mario Santisteban Galarza*

The applicability of the Artificial Intelligence Act to the health sector and specialities regarding its compliance, *by Iñigo De Miguel Beriain*

The applicability of the Artificial Intelligence Act to the field of public administration and public services and special features regarding compliance: Special attention to Annex III and administrative action and particularities of compliance, *by Gal-la Barrachina Navarro and Andrés Boix Palop*

Large Artificial Intelligence platforms and systems for political influence: the intersection between the "Digital Services Act" and the Artificial Intelligence Act from a risk perspective, *by Rosa Cernada Badía*

General regime for high-risk Artificial Intelligence systems

The application of harmonised standards and common specifications in the field of Artificial Intelligence (Articles 40 and 41 AIA), *by Vicente Álvarez García*

Conformity assessment in the design and production of Artificial Intelligence-based systems in the context of the "New Legislative Framework", *by Adrián Palma Ortigosa*

General regime of obligations for providers and deployers in the Artificial Intelligence Act, *by Adrián Palma Ortigosa*

Subjects and actors in conformity assessments (notified bodies), *by Ignacio Alamillo Domingo*

Obligations of providers and deployers of high-risk systems

The fundamental rights impact assessment by deployers of Artificial Intelligence systems in the Regulation, *by Eduard Chaveli Donet*

Risk management systems as a specific obligation for high-risk Artificial Intelligence systems in Article 9 of the Regulation, *by Pere Simón Castellano*

Data and data governance and connections with data protection principles in Article 10 of the Artificial Intelligence Act, *by María Loza Corera*

Quality management systems, technical documentation and documentation keeping in the Regulation, *by Francisca Ramón Fernández*

The obligation to keep records of high-risk systems in the Artificial Intelligence Act, *by Wilma Arellano Toledo and Antonio Merchán Murillo*

Transparency and provision of information to deployers in Article 13 of the Artificial Intelligence Act, *by Estrella David Gutiérrez*

Human oversight or monitoring in Article 14 of the Artificial Intelligence Act: a mere mandatory requirement for high-risk systems? *by Guillermo Lazcoz Moratinos*

Accuracy and robustness of high-risk Artificial Intelligence systems in Article 15 of the Artificial Intelligence Act, *by Ana Aba Catoira*

Cybersecurity in high-risk Artificial Intelligence systems in Article 15 of the Artificial Intelligence Act, *by Marco Emilio Sánchez Acevedo*

Post-market monitoring in high-risk AI in the Artificial Intelligence Act. Description, measures and use cases, *by Idoia and Miguel Ángel Liébanas*

General-purpose Artificial Intelligence, non-high-risk systems and Article 50 systems

General-purpose Artificial Intelligence, foundational models (and "GPT Chat") in the Artificial Intelligence Act, *by José Antonio Castillo*

Codes of conduct, seals or certifications for Artificial Intelligence systems that are not high-risk (Article 95 of the Regulation, *by Lorenzo Cotino Hueso*

Article 50 of the AI Act and the transparency obligations for providers and deployers of certain Artificial Intelligence systems, *by Agustí Cerrillo i Martínez*

Sandbox, governance, oversight, sanctions, rights and confidentiality in the Regulation

Sandbox, controlled spaces and real-world testing of Artificial Intelligence systems in the Regulation. Measures for SMEs, start-ups and micro-enterprises, *by Lorenzo Cotino Hueso*

Governance and oversight of the Artificial Intelligence Act: market surveillance authorities, Commission and the various entities, *by Juan Carlos Hernández Peña*

The sanctioning regime in the Artificial Intelligence Act, *by F. Javier Sempere*

Right to lodge a complaint and right to an explanation. Means of redress for individuals in the Artificial Intelligence Act, *by Aurelio López-Tarruella Martinez*

Access to documents and confidentiality in the Artificial Intelligence Act, *by Gabriele Vestri*

PRESENTATION AND INTRODUCTION

By Lorenzo Cotino Hueso and Diana-Urania Galetta

Artificial Intelligence (AI) has already proven to be one of the most influential and disruptive technologies of our era, as a result of its remarkable ability to swiftly transform productive sectors, from medicine to security, and also, among many other revolutions, due to its possible integration into digital objects, products or services. In fact, AI solutions have been implemented in the private sector already for decades[1], largely passing unnoticed. Virtual voice assistants, including Alexa, Siri, and Cortana, are among the daily applications of Artificial Intelligence, which are used by millions of individuals worldwide. Many people have also smart speakers installed in their residences, which use natural language processing (NLP) and machine learning technology to interpret and respond to spoken commands in order to gain a more comprehensive understanding of the information being conveyed.

Social networks also employ AI: the content that users peruse on a daily basis is structured to be measurable. In accordance with their preferences and the frequency with which they access specific photos, videos, or activities, Artificial Intelligence is responsible for their selection. Similarly, via Artificial Intelligence (AI), social networks can generate more precise recommendations regarding advertisements and friendships. Streaming services like Netflix and Spotify implement AI and machine learning technologies, as well, to personalise the content they provide in accordance with the preferences and interests of their users. The process is similar for email providers: they employ Machine Learning technology to differentiate between important messages and those that are deemed spam. Additionally, they now often implement a "Smart Reply" feature, which involves the prediction of words and sentences to facilitate the composition of messages in a more agile manner.

Therefore, at least in the European Union (EU) context, the need for a robust and coherent general regulatory framework to accompany and guide the development of this disruptive technology has been perceived for years. More in general, the European Union has long been committed to the goal of achieving its own "digital sovereignty", both internally and externally, in order

[1] See B. Rashid, A. K. Kausik, *AI revolutionizing industries worldwide: A comprehensive overview of its diverse applications*, in *Hybrid Advances*, 7/2024, p. 1 ss.

to strengthen its geopolitical position by introducing a regulatory framework that can serve as a model, and be exportable, externally as well.

This strategy of attempting to export a regulatory model has indeed been known for some time[2]. The European Union, for better or for worse, has been the pioneer in establishing this general regulation om AI both for itself and to try to influence the rest of the world with what has been called the *Brussels effect,* something that, to some extent, was achieved already with the approval of the General Data Protection Regulation (hereafter GDPR).

The EU Artificial Intelligence Act (AIA)[3] has undergone a lengthy and costly legislative procedure[4], in which the European Commission's proposal of 2021 and the EU Council's position of December 2022 are currently particularly noteworthy[5]. In addition, the amendments of June 2023 by the Parliament are very important[6].

The AIA has integrated the regulation of AI into the "New Legislative Framework" model, harmonisation standards, and the area of product safety and assurance. This is the framework that establishes a shared foundation for the marketing, evaluation, and surveillance of products in the European Union[7]. This model is not well-known to the majority of legal professionals. The Commission, Parliament, and Council were involved in the initiative to address AI in the EU over seven years ago. An AI that is ethically designed

[2] On this topic see D.U. Galetta, *El nuevo protagonismo de la Unión europea como organización supranacional y como actor en el tablero mundial*, in *Actas del XVII Congreso de la Asociación Española de Profesores de Derecho Administrativo (Sevilla 26 a 28 de enero de 2023). 20 Años de La Ley General de Subvenciones*, Instituto Nacional de Administración Pública, Madrid, 2023, p. 499 ss.

[3] Its proper name is Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

[4] See at: https://oeil.secure.europarl.europa.eu/oeil/en/procedure-file?reference=2021/0106(COD). See also at European Parliament Legislative Train Schedule: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence.

[5] At https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf.

[6] At https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html.

[7] The three legal texts that make up the New Legislative Framework are: Regulation (EC) No 765/2008 of the European Parliament and of the Council setting out the requirements for accreditation and market surveillance of products; Decision No 768/2008/EC of the European Parliament and of the Council on a common framework for the marketing of products and; Regulation (EU) 2019/1020 of the European Parliament and of the Council on market surveillance and product conformity.

to respect democratic principles and liberties and is “made in Europe” is the objective[8].

The European Commission’s proposal was designed to promote investment and innovation in the field of AI in Europe, as well as to guarantee the safety of AI systems and their respect for citizens’ rights when they are implemented in the EU and introduced to the European market. The provisional agreement of the Council and the European Parliament on 8 December 2023, the approval by the Internal Market and Civil Liberties Committees on 13 February 2024, the Resolution of the European Parliament on 13 March 2024 (*corrigendum* of 16 April 2024), and the final approval by the Council of the European Union on 21 May 2024 have all been significant milestones in the process of the proposal’s debate, study, and amendment, which has followed the ordinary legislative procedure[9].

Using a risk-based approach, the final text of the Regulation seeks to harmonise the rules on Artificial Intelligence, establishing a framework of varying obligations and requirements based on the level of risk associated with the applicable AI technology and its specific use. For instance, the harmonising approach necessitates the establishment of *ad hoc* controls for the implementation of technical standards and the enhancement of obligations for Artificial Intelligence systems classified as high risk.

Although the AIA has entered into force in 2024, it will certainly have a very staggered application and enforceability, taking up to six years. In any event, it is feasible to anticipate a huge impact on the market and the society. The establishment and operation of European and National AI supervisory offices and authorities, as well as the prohibition of specific technologies and applications, will occur in the near future. The new AI Office was established already in February 2024.

We are already witnessing bans on specific technologies and uses, the establishment and activity of European and national AI supervisory offices and authorities. Already in February 2024 the new AI Office has been established. We will further witness the initiation and resolution of sanctioning procedures, the approval of technical standards and the activity of certification bodies, the proliferation of specific risk management systems, the emergence

[8] For an exhaustive analysis of the steps and policies of the EU in this area until 2019, see L. Cotino Hueso, “Ética en el diseño para el desarrollo de una inteligencia artificial, robótica y big data confiables y su utilidad desde el derecho”, in *Revista Catalana de Derecho Público* n.º 58 (June 2019). http://revistes.eapc.gencat.cat/index.php/rcdp/issue/view/n58 http://dx.doi.org/10.2436/rcdp.i58.2019.3303

[9] At https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf.

of sectoral voluntary codes of conduct, among many other issues that will be a reflection of the legal framework that is the subject of study in this commentary.

This book provides a detailed, systematic, and comprehensive understanding of the new AIA.

As editors of the book we believe that its approach is genuinely innovative in that we deliberately rejected the traditional model of commentaries on legislative acts, which are typically published in the order in which the articles appear in the legislative act. We could have provided a concise commentary for each article; however, this classification would not have made a substantial contribution in terms of content, organisation and structure. Therefore, we have endeavoured to establish a classification based on thematic blocks in order to try and offer a comprehensive and global response to the theoretical and practical implications of the approval of a Regulation that contains 180 recitals, 112 articles, and 13 annexes. In contrast to the GDPR's 60,000 words, the AIA contains approximately 108,000 words.

The 38 chapters that make up the collective work we present here include 34 authors, 30 of whom hold PhDs and 29 of whom are lecturers. The current (English language) edition of this book is a translation and adaptation of the Spanish original one, for which we had selected only the leading researcher in Spain in each of the subjects.

In spite of the fact that many people currently claim to be expert in both law and AI, the truth is that only a small number of scholars have dedicated decades to the study of digital law and have already dedicated a few years to the study of AI and its implications, in particular. Certainly, this is the case with the authors of this commentary. Some of them share authorship and others have been responsible for writing two or more chapters. We are also fortunate to have the participation of leading international figures in the field, such as Corvalán, Mantelero and Ziller.

The editors (*Lorenzo Cotino Hueso* and *Diana-Urania Galetta* for this English language edition of the Systematic Commentary to the EU AIA) are deeply grateful for the generous contributions of more than thirty individuals to this book. They are aware that the selection of experts, the distribution and delimitation of their work, the discipline of authors to adhere to the time and standards of the work, and, of course, the review of such work are not always simple or pleasant.

With regard to the thematic organization of the book, the first part is dedicated to the study of AIA and its global contextualisation, from Latin America and Europe. In this section we count on the significant contributions of Alessandro Mantelero, Jacques Ziller, Juan Gustavo Corvalán and María Victoria Carro, all of them outstanding voices at the intersection between Law and AI. The second part of the book resolves terminological issues (what the AIA regulates, exclusions, and what is meant by AI), the territorial scope and reach of the AIA, as well as its relationship with data protection. The third section includes an analysis of AI that is prohibited or unacceptable for the AIA. The fourth section then analyses high-risk AI systems, identifying and analysing the most controversial or highly contentious areas. A fifth section is dedicated exclusively to the general framework that applies to high-risk AI systems, including the application of harmonised standards and the investigation of conformity assessment models and notified bodies. The sixth section introduces the comprehensive set of obligations that are specific to providers and deployers of high-risk AI systems. The regulation of systems that have not been classified as high-risk AI, general-purpose AI, and Article 50 AIA systems is the focus of Part seven. Lastly, the eighth section is designated for the examination of governance and compliance oversight mechanisms, the sanctioning regime, the potential for the establishment of regulatory sandboxes and controlled testing areas, the right to lodge a complaint and to obtain an explanation of the AI, and the possibility for access to documentation and confidentiality in the AIA.

In summary, this systematic commentary is intended to serve as a valuable resource for comprehending and implementing the AIA. Through the comprehensive thematic structure of a complex harmonisation technique and the collaboration of prominent experts in the field, it provides a detailed and exhaustive perspective on the numerous regulatory, technical, and ethical issues that this regulation raises. The objective is to not only offer practical and theoretical guidance to academicians, practitioners, and policymakers, but also to engage in a broader discussion regarding the global impact and future of AI in Europe. The depth and rigour with which each topic is examined are indicative of the authors' commitment to academic excellence and practical relevance. Subsequently, we, the Editors, believe that this work will serve as not only a reference book, but also an indispensable resource in the field of AI law.

Finally, we wish to express gratitude and acknowledge to those who provided assistance with this project. Namely, the coordination and publication of this systematic commentary is a result of the MICINN Project "Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas" 2023-2025 (PID2022-136439OB-I00) financed by MCIN/AEI/10.13039/501100011033/ which financed the publication of its original Spanish edition.

As for this English version of the systematic commentary to the EU

AIA, translation into English, text adaptation etc. were funded by the "Algoritmic Law" Project (Prometeo/2021/009, 2021-24 Generalitat Valenciana).

This book is also related to (and a product of) projects such as "Algorithmic Decisions and the Law: Opening the Black Box" (TED2021-131472A-I00) and "Transición digital de las Administraciones públicas e inteligencia artificial " (TED2021-132191B-I00) of the Recovery, Transformation and Resilience Plan. Generalitat Valenciana CIAEST/2022/1. Also, of the Digital Rights Agreement-SEDIA Scope 5 (2023/C046/00228673) and Scope 6 (2023/C046/00229475).

It is a source of pride for the editors to have had comprehensive recognition and research support from the various institutions mentioned, for many years now.

To finish, a special thank you to Mercedes López de la Peña de Pablo, specialist in law and Artificial Intelligence, who did a fantastic job of translating and revising the content, starting with an automated translation from Spanish to English.

Artificial Intelligence Act and its global contextualisation, from Ibero-America and Europe

THE ARTIFICIAL INTELLIGENCE ACT: THE EUROPEAN LEGISLATOR'S RESPONSE TO THE CHALLENGES OF ARTIFICIAL INTELLIGENCE

Alessandro Mantelero

Senior Lecturer in Civil Law at the Politecnico di Torino and holder of the Jean Monnet Chair in Mediterranean Digital Societies and Law

I. Introduction

What is the vision of the European legislator in the regulation of Artificial Intelligence? What is the relevance of adopting a risk-centred paradigm? How does this paradigm intersect with the fundamental rights dimension? These are the main questions that a first review of the AIA aims to answer, highlighting the need for an interdisciplinary approach, also looking at international and other countries' scenarios, to fully understand the dynamics that inspired the European legislator and that will guide the implementation of the AIA.

In order to contribute, with a first brief reflection on the finally adopted text of the AIA, to the growing legal debate on AI, the following pages will examine the core of this regulatory framework with a focus on legal policy options.

Due to the nature and relative space of this reflection, we will not give an account of the various issues that have fueled and continue to fuel the doctrinal debate regarding the different aspects of the relationship between AI, law and society, both in our legal system and in others, leaving the reader the opportunity to delve deeper into these profiles in the already extensive bibliography available.

With regard to the examination of AIA, in the following discussion we have chosen to give priority to answering three main research questions: (i) What is the European legislator's vision in regulating AI? (ii) What is the normative relevance of adopting a risk-centred paradigm? (iii) How does the so-called risk-based model relate to the fundamental rights dimension?

II. The European perspective

In 1968, Stanley Kubrick staged 2001: A Space Odyssey, in which an Artificial Intelligence was concerned with the welfare of human beings, but then turned malevolent and gave rise to an iconic confrontation between human and machine will. It was certainly not the first time that automatons and machine intelligence had been fantasised about, but it was no coincidence that the film was released in the same years when, alongside Alan Westin's seminal work from 1967, a series of critical books on the role of computers in the new digital society were being published.[1]

These were the years when the potential of ICTs were already becoming apparent, and the foundations were being laid, even if the tools were still inadequate to develop their full potential. There was already talk of AI, expert systems and automation algorithms, but there was a lack of huge amounts of digitised data and computers capable of processing it. As with steam, the telegraph and many other inventions, the ideas were there, but their application was in its infancy.

However, the very vision of the potential of information technology, even then, led to thinking in terms of social impact with a clear tension between the new utility brought by the technologies and the relative risk. In the 20th century, the experience of wars, the uncertainty of scientific paradigms, and the perceived weakness of human beings shattered the uncritical faith in progress that had characterized previous centuries. Progress was thus joined by *hybris*, in the challenge of generating something fascinating and terrible (as had been the case with the atom), which counterposed the positive vision of the American counterculture to questions about the future of a world characterised by automated processes.

It could be argued that all this refers to the past and has little legal relevance in relation to the commentary on the AIA that is the subject of these brief notes. However, it would be useful to start again from Westin, to recall how the jurist cannot reflect on what the rules are without taking into account the strong forces that characterise society and generate the context to which the legal rules, mere instruments, are called upon to provide one of the possible responses.

Thus, reaching the present day, one cannot understand the AIA and the tenor of its provisions without keeping in mind the US and Chinese dominance of AI markets, the risky move by Open AI (read Microsoft) to bring an immature technology like ChatGPT to the market, or the systemic use by totalitarian states of biometric and social control tools. Listing the categories of prohibited uses of AI contained in the AIA, discussing the rules on gener-

[1] See, for example, Miller, *The Assault on Privacy – Computers, Data Banks, Dossiers*, Ann Arbor, 1971; Brenton, *The Privacy Invaders. Coward-McCann*, New York, 1964; Packard, *The Naked Society*, New York, 1964.

al-purpose models (GPAI) – in particular large generative models -, addressing the issue of impact assessment, would be incomprehensible exercises if they were only considered from the perspective of abstract legal categories.

From this perspective, we must first place AIA in its relevant geopolitical context. Indeed, this legislation does not come out of nowhere, nor does it arise solely from needs related to the potential impacts of AI, but is part of a broader EU design for a digital society. When the new European Commission's strategic plan was presented in 2019, digital regulation was positioned as a key element of EU legislation for the period 2019-2023. At that time, only a few digital society regulations existed, mainly Directive 95/46/EC on personal data, its daughter directive on e-privacy, Directive 2000/31/EC on e-commerce (central especially in the area of suppliers' liability) and the Public Sector Information Directive (Directive 2013/37/EU). Today, there are dozens of regulations adopted or about to be adopted at European level.[2]

There is, therefore, a regulatory policy strategy that goes far beyond AIA and that needs to be understood in order to properly assess its scope. There are several guidelines that have led the European legislator to make such an intense, perhaps even excessive, regulatory effort during the legislature that will end in 2024.

First, of course, there are the changes in the structure of the digital society. After the distributed computing of the 1980s and the arrival of the Internet in the 1990s, from which the first data and e-commerce regulations emerged, the explosion of sensors (read IoT) and computing power (read cloud computing) have paved the way for AI, but also for new threats on the cybersecurity and social impact front. At the same time, the concentration that has characterised the last decades of the digital economy, together with the overcoming of the distinction between the *online* and *offline* worlds, has left only the memory of an environment of small players to protect against the legal risk of their pioneering investments in the digital sector, and has demanded more effective responses to the dominance of global platforms.[3]

With regard to these first factors, AIA is a necessary and coherent response, both because it is precisely the technological paradigm shift (abundance of data and computing power, omnipresence of technology and data collection, widespread diffusion of human-machine interaction systems) that has enabled the latest AI revolution, and above all because this revolution is based on phenomena of concentration of information and market power. In

[2] For a map of relevant EU legislation, see e.g. https://www.bruegel.org/sites/default/files/2023-11/Bruegel_factsheet.pdf.

[3] See the Digital Markets Act and the Digital Services Act.

fact, it is no coincidence that the most advanced and critical applications of AI, in the field of GPAI, are the prerogative of an extremely limited number of operators on a global scale, from which derives a strong power to condition the market and the geopolitical scenario, given their prevalent location in the US and China.

It is precisely the geopolitical scenario that is the second soul of the EU's regulatory wave in the digital domain. Here the focus is on the chronic weakness of Europe's industrial sector compared to its Asian and North American competitors. From raw materials to platforms, the EU has failed to gain technological dominance on the global stage. Moreover, due to aggressive takeover policies of the most innovative companies by the big players, Europe is now largely a land of colonisation for foreign digital multinationals. At the same time, the gigantism of these multinationals and their weight in conditioning digital society has recently led them to act as quasi-state realities, not only autonomously and self-referentially defining policies in digitally mediated social relations (think, for example, of the cultural dimension of content moderation policies), but also often (from smart cities to pandemics) exercising functions of the state.[4]

In this context, therefore, there is a clear need for the EU to provide itself with legislation that regulates the digital sector across a broad spectrum. Since it cannot in fact use the so-called *bully pulpit*, as Reidenberg referred to,[5] to be able to indirectly condition technology producers,[6] has only the exogenous recourse of binding regulation to protect European interests.

The establishment of binding rules is another element that characterises AIA. This position, in terms of legal regulation, is highlighted not only by the recourse to legislative intervention instead of the use of *soft law* typical of AI-producing countries, but also by the specific provisions on the territorial effectiveness of the AIA. Indeed, the AIA provides for its applicability to providers of AI systems available in the EU regardless of the establishment

[4] See, for example, on the subject Goodman, Powles, *Urbanism Under Google: Lessons from Sidewalk Toronto*, in *Fordham Law Review*, 2019, 88 (2), 457 et seq.

[5] See Reidenberg, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, in *Texas Law Review*, 1998, 76 (3), 553, 581 et seq. ("Government can use the bully pulpit approach to threaten and cajole industry to develop technical rules [...] The government's bully pulpit resulted in a flexible mechanism that can provide an information policy rule customized by network participants rather than an immutable architectural rule").

[6] See in this regard, The White House, *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*, 30 October 2023, https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/.

of the provider, including the case where the providers and *deployers* of AI systems are located outside the EU, but the output generated by the AI systems is used in the EU.

However, placing European interests at the heart of the regulatory approach obviously requires defining what they are, or rather prioritising those that constitute the fundamental purposes of the Union. In this respect, a comparison of the legislative drafting process of the AIA with that of the GDPR reveals significant and indicative differences. The leadership of the Directorate-General for Justice, focused on freedom, security and justice, which had characterised the drafting of the GDPR, is here replaced by that of the Directorate-General for the Internal Market, Industry, Entrepreneurship and SMEs, and the focus of the AIA is clearly that of an industrial security regulation, as often presented by the Commission.

In this way, a strong hiatus is evident between the way in which AI risks were metabolised in the legal political discourse, focusing on ethical issues[7] and fundamental rights, although not infrequently with an unfortunate confusion between the two agendas, and the Commission's own vision aimed primarily at industrial safety, with a broad focus on risk management in terms of conformity assessment and with significant weight given to standards.

The Commission's proposal included references to the protection of fundamental rights in relation to the potential impacts of AI, but without specifying them. This contrasted with a public debate that suggested that the risks of AI, for example, had less to do with the harm that the collaborative robot may do to the worker and more to do with the potential discrimination and misinformation that algorithms are introducing into society. It was only after the outcome of the parliamentary debate, also thanks to the support of international academia,[8] that the AIA was made to contain more detail on the impact on fundamental rights.

However, this connotation of the regulatory proposal, briefly outlined here, highlights the European legislator's objective, which is not primarily to protect fundamental rights, but to encourage the development of AI in a context of industrial weakness in the sector in Europe. Hence the focus on

[7] Consider the initiatives of the European Data Protection Supervisor, as well as the more questionable contribution of the High Level Expert Group on Artificial Intelligence, of which there is a trace in one of the recitals of the Regulation.

[8] Brussels Privacy Hub, *More than 150 university professors from all over Europe and beyond are calling on the European institutions to include a fundamental rights impact assessment in the future regulation on Artificial Intelligence*, 12 September 2023, https://brusselsprivacyhub.com/2023/09/12/brussels-privacy-hub-and-other-academic-institutions-ask-to-approve-a-fundamental-rights-impact-assessment-in-the-eu-artificial-intelligence-act/

industrial safety and, above all, the balance chosen in risk management, which will be analysed in more detail below. Here, in general terms, it is sufficient to note how the industrial policy choices led to a more risk-accepting perspective, different from the more marked risk aversion observed in the GDPR.[9]

It is also worth noting that the intense activity of the European legislator in digital matters that has characterised recent years raises a number of systemic issues that also affect AIA. First, the fragmentation of the different initiatives in terms of promoters leads to texts being drafted more in silos than in a systematic way. As already pointed out by regulators such as the EDPB and the EDPS, the drive to develop a new regulatory framework, induced by the above-mentioned reasons, produced many rather lengthy and complex texts in a rather short period of time, without any fine-tuning of coordination between them.

Secondly, this extensive regulatory effort affected the timing of the various approval processes, with the result that some areas are only partially regulated: one example is the non-approval of the complementary pillar of the AIA, i.e., the directive on liability related to the use of AI.

On the other hand, the approach adopted by the European legislator is characterised by considerable pragmatism, seeking a regulation focused on *ex-ante* remedies, in terms of risk management and product *by-design* approach rather than *ex-post* compensatory measures. Increasingly, this is leading to liability rules as the cornerstone of regulation aimed at preventing the risks of complex systems. Indeed, the traditional recourse to tort liability is ill-suited to an environment characterised by technological complexity, global operators with large financial resources, the pulverisation of damages, and, in order to foster trust in new technologies (so-called trusted AI), the need to ensure safe technological environments rather than compensation remedies in case of disastrous consequences.

However, the lack of a system of lock-in rules on liability in AI -given the issues regarding its distribution both with respect to the various components of AI systems and to human-machine interaction- points to a lack of coordination in the European approach. Other legislators, think of the Brazilian proposals on AI, have more adequately combined risk management, with penalties for non-compliance, and liability for harm caused by AI.

The decision to keep the two profiles separate was therefore unfortunate, as was the decision to address them with two different legislative instruments and, moreover, to promote a parallel update of the product liability framework in general. Developing an *ex ante* protection model, focused on risk analysis, without then developing an adequate framework for the residual hy-

[9] See Art. 35 GDPR.

potheses in which poor or deficient risk management causes damage, ends up undermining the overall impact of the regulatory intervention derived from the AIA, which is thus an unfinished work in a broad view of the regulation of AI. Nor is it possible to argue here, unlike in the case of personal data protection, that compensation profiles hold limited relevance. The delegation of critical infrastructure management functions to AI systems, both functionally and socially, suggests broader scenarios of potential harm.

III. The modulation of the so-called risk-based approach in first-generation legislation

Secondly, an approach focusing on the classification of high-risk cases was preferred in order to make it easier for operators to know from the outset whether or not they are subject to the new rules. This choice, aimed at apparent simplification, turned out to be inherently complex due to the difficulty of providing a precise definition of high-risk systems and the evolving uses of AI. Hence the corrective measures, such as the possibility of exemptions[10] and future amendments to Annex III,[11] with an overall framework which, rather than simplifying, threatens to make the landscape of industrial exploitation of AI tortuous and open to litigation.

The criteria according to which high-risk cases are dealt with in terms of mitigation strategies also deserves attention. According to the industrial risk model, it is considered that AI development can be justified[12] even if it carries high risks. In consequence, the criteria of acceptability of residual risk, which

[10] See art. 6.4, AIA ("A provider who considers that an AI system referred to in Annex III is not high-risk shall document its assessment before that system is placed on the market or put into service. Such provider shall be subject to the registration obligation set out in Article 49(2). Upon request of national competent authorities, the provider shall provide the documentation of the assessment.").

[11] See art. 7, AIA, which gives the Commission the power to "The Commission is empowered to adopt delegated acts in accordance with Article 97 to amend Annex III by adding or modifying use-cases of high-risk AI systems where both of the following conditions are fulfilled: (a) the AI systems are intended to be used in any of the areas listed in Annex III; (b) the AI systems pose a risk of harm to health and safety, or an adverse impact on fundamental rights, and that risk is equivalent to, or greater than, the risk of harm or of adverse impact posed by the high-risk AI systems already referred to in Annex III".

[12] See also the express reference to the benefits of AI in art. 7, AIA ("When assessing the condition under paragraph 1, point (b), the Commission shall take into account the following criteria: [...] "the magnitude and likelihood of benefit of the deployment of the AI system for individuals, groups, or society at large, including possible improvements in product safety"").

does not necessarily need to be high, but only justified by other overriding interests.

This acceptability criterion is developed through the risk assessment provided for in the AIA, i.e. the conformity assessment provided for in Article 43.[13] However, it should be noted that one component of this assessment is also the assessment of the impact on fundamental rights and freedoms, for which acceptability based on an indiscriminate comparison of the interests at stake seems to be ruled out.

The European and the Member States' legal systems provide a level of protection to fundamental rights that precludes their compression due to conflicting interests, social acceptability, and acceptable residual risk[14]. A necessary and proportionate compression of fundamental rights can only be justified in the event of a conflict with interests considered equal or superior by the legal system.

Finally, there are cases in which the European legislator considered certain uses of AI to be unacceptable precisely because of their stark contrast with fundamental rights and the principles of EU law. These are those identified in Article 5 of the AIA, including manipulative techniques, so-called *social credit scoring* and certain invasive uses of biometric technologies. In this respect, a broad debate took place, with the participation of civil society, on the identification of prohibited uses and on the exceptions (quite articulated, especially as regards the use of biometric identification) that were added throughout the legislative process.

Rather than relying on the often mentioned pyramid of risks (many unregulated AI systems, some subject to limited obligations, a few with high risk subject to compliance assessment, very few prohibited),[15] the general pattern that shows up in the way AI-related risks are dealt with in the regulatory framework should be reconstructed using the three different ways of evaluating risk.

In this respect, a distinction should be made between a near technology assessment, a conformity assessment and a fundamental rights impact assessment. The first is the exercise elaborated in Art. 5 of the AIA to define the prohibited categories. This is an *ex ante* assessment formulated in the ab-

[13] See also Articles 9 and 17.

[14] In some circumstances, residual risks cannot be excluded, but this implies ex-post additional measures, such as compensation, rather than their acceptability.

[15] This pyramid model actually provides little insight into legislative policy options and is, above all, functional to a narrative that wants to emphasise minimalist intervention, limited to the most serious cases, by the European legislator, underlining the innovation-friendly orientation of the AI Regulation.

stract on new uses of technology whose regulatory acceptability is assessed in terms of their impact on the founding principles of EU law. An example is the use of subliminal technologies intended to manipulate the individual will of "an AI system that uses subliminal techniques that transcend a person's consciousness or deliberately manipulative or deceptive techniques with the aim or effect of substantially altering the behaviour of a person or a group of persons, appreciably impairing their ability to make an informed decision and causing a person to take a decision that they would not otherwise have taken, in a way that causes, or is likely to cause, significant harm to that person, another person or a group of persons". Within the same type of assessment is also the list of high-risk uses in Annex III, where, again, AI systems are considered in terms of categories of use, regardless of their specific configuration and contextual use.[16]

In this respect, while for the possible variation of the categories prohibited, due to technological evolution and the sociotechnical context, it is planned to resort to subsequent amendments to Article 5 of the AIA, the evaluation of the high-risk systems are left for future work by the European Commission. The latter option, while allowing the Commission to act within the limits defined by Article 7 of the AIA, nevertheless implies giving it the possibility of amending the subject matter of the legislation, which seems peculiar, given the institutional nature of the Commission and the legitimacy of the Union's legislative process.

Conformity assessment, on the other hand, is of a different nature. Whether it is based on the procedures of Annex VII (Conformity based on quality management system assessment and technical documentation assessment) or Annex VI (Conformity assessment procedure based on internal control), depending on whether or not the use of high-risk biometric technologies as defined in Annex III is involved,[17] always requires the implementation of a quality management system in accordance with Article 17 of the AIA, of which the risk management system in accordance with Article 9 is a central component, and which also includes the assessment of the impact on fundamental rights.

Conformity assessment, in contrast to technology assessment, is an as-

[16] Reference is made, for example, in the field of education to "AI systems intended to be used to determine access or admission or to assign natural persons to educational and vocational training institutions at all levels", where there are various possibilities for the configuration of such systems depending on the parameters used and the thresholds adopted, as well as different implications of application depending on the specific socio-cultural context of use.

[17] See Art. 43(1) and (2) AI Act.

sessment focused on a specific use of AI, characterised by specific and differentiating functionalities, although it can be used in different scenarios.[18] This assessment focuses on industrial risk in traditional terms of harm to the physical integrity and safety of the product/service, but also includes risks in terms of harm to fundamental rights.[19] In this respect, the approach of the European legislator is to leave this conformity assessment to the adoption of standards.[20]

It is worth noting how recourse to the standardisation process for conformity assessment is coherent with the practice of industrial and product risk management in terms of safety (including the physical safety of humans interacting with machines, here AI), but seems inadequate with regard to the fundamental rights impact assessment component. As regards the latter, not only the opacity of the standardisation systems, but also the lack of involvement of fundamental rights experts is a first critical issue, stigmatised even in the draft standardisation request submitted by the Commission to CEN-CENELEC, whose lack of expertise on fundamental rights is explicitly admitted.[21]

Beyond the structural problems of the standardisation system, there is a more important methodological objection to the difficulty of using standards to assess the impact on fundamental rights. Indeed, standards, by their very nature, are usable in the presence of processes characterised by constant and repetitive dynamics, so that it is possible to define a standard in railway construction, since the variables of speed, weight, gradient, etc. move within constant ranges with respect to a train running activity that happens to have uniform characteristics regardless of the different layouts.

This uniformity cannot be seen in the context of the impact of AI on fundamental rights, where the same AI application can have significantly dif-

[18] In line with the hypothesis set out above, see footnote 15, the assessment of compliance will refer to a specific AI application which, based on parameters relating to the student's grade in a given time frame, performance in different subjects, age, time series used in the training phase and many other parameters, will be able to assess admission to a given degree programme. Therefore, it will not be a type of AI application, but a specific product with its own design and training options, although it will be susceptible to be applied in different contexts in terms of demographic variables, type of career, etc.

[19] See art. 9.2.a, AIA ("identification and analysis of the known and the reasonably foreseeable risks that the high-risk AI system can pose to health, safety or fundamental rights when the high-risk AI system is used in accordance with its intended purpose").

[20] See art. 40, AIA (Harmonised standards and standardisation deliverables).

[21] See European Commission, *Draft standardisation request to the European Standardisation Organisations in support of safe and trustworthy Artificial Intelligence*, 5 December 2022, https://ec.europa.eu/docsroom/documents/52376?locale=en.

ferent impacts due to the characteristics of the technologies used, the context of use and the actors involved. If we consider, for example, AI-based video surveillance systems, in terms of their impact on fundamental rights, there are different scenarios depending on whether they are used in public or private spaces, whether minors or other vulnerable persons are present in the latter, whether real-time monitoring functionalities are implemented or not, whether they are used in contexts characterised by high levels of criminality with the aim of fighting crime, and depending on many other factors that could be added due to the variety of possible scenarios.

It is therefore clear how the variability of the contextual dimension of fundamental rights impact assessment cannot be reconciled with an idea of standardisation, if by standardisation we mean the possibility of defining a precise and uniform procedure, composed of specific stages of shaping the technology according to predefined patterns. On the other hand, the conclusion may be different if standards are understood as methodological rules, i.e. not the definition of a specific process, but rather as a general methodological framework for risk management in the case of fundamental rights, for example with regard to the central question of the impact assessment criteria needed to compare different design options in AI development.[22]

Finally, following the debate in the European Parliament, a specific obligation for a Fundamental Rights Impact Assessment (FRIA)[23] by those responsible for the deployment of AI systems was introduced in the AIA. This assessment can be partly developed by the AI provider on the basis of possible scenarios of use, as is the case for conformity assessment, but must also take into account the concrete application of AI in the specific case. This is in line with human rights impact assessment and data protection impact assessment processes, which are based on contextual assessments of the potential harm to the rights and freedoms at stake.

According to the general theory, with reference to the distribution of risks and corresponding responsibilities, the assessment of the impact on fundamental rights is thus combined with the assessment of compliance, transferring to those responsible for the deployment of AI systems part of the burden of managing the potential negative consequences of AI linked to the specific operational context of use, in respect of which those responsible parties have greater margins of control or, at least, of actual risk assessment.

[22] For an example of this approach, see Mantelero, *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI*, The Hague, 2022, chapter 2, https://doi.org/10.1007/978-94-6265-531-7 (open access).

[23] See Art. 27, AIA.

Finally, unlike conformity assessment, no standardisation processes are foreseen for fundamental rights impact assessment. This is in line with previous experiences with fundamental rights impact assessment and data protection impact assessment, where best practice risk assessment models have emerged.

Despite the tripartite assessment process in the AIA and its implementation, the approach lacks uniformity, as evidenced by the absence of common risk assessment parameters and methodologies for evaluating the impact on fundamental rights in the context of AI. Thus, while on the one hand risk is defined in general terms in Article 3 as a combination of likelihood and severity of the harmful event, Article 7 lists a number of additional parameters in relation to the assessment of the technology. On the other hand, specific indications for conformity assessment are missing, and, in a questionable attempt at simplification during the trialogues, references to the parameters to be taken into account in the assessment were deleted even in the final text of Article 27 on the impact on fundamental rights, as opposed to the more precise Parliamentary proposal. Therefore, conducting a methodological reflection on how to conduct the impact assessment, particularly in relation to fundamental rights, is crucial for the successful implementation of the AIA.

With respect to this basic structure of the AIA, two blocks of provisions should be considered below, relating respectively to the transparency obligations foreseen for non-high risk systems and to the provisions added in the final drafting phase of the Regulation to address concerns raised by general-purpose model-based AI systems (GPAI), which were made known to the general public especially after the publication of ChatGPT.

As regards the first set of rules, the AIA adopts what the Council of Europe already indicated in its guidelines on AI and personal data protection about the obligation to make the end-user aware of the fact that he or she is interacting with an AI system. This is an obligation justified by the ability of AI to emulate various human behaviours in human-machine interaction. Additional specific transparency obligations are also imposed on both producers and deployers of AI systems in relation to AI's ability to generate synthetic content, especially considering the critical issues that this may when it comes to altering reality with significant social repercussions (for example, fake news).[24]

More complex is the discourse in relation to general-purpose model-based AI (GPAI), where the haste due to the emergence of the problem at the final stage of the legislative process and the opposing positions of some

[24] See Art. 50, AIA.

governments to an incisive regulation of this important aspect have led to the outline of a regulation that could be defined as minimalist.

On this point, reflection should go far beyond the limited considerations set out in these pages, raising questions that are at the root of the problem of regulating the technology and that have to do with the well-known Collingridge dilemma, where the GPAI is a technology still in its infancy, not by chance afflicted by several unresolved operational problems and also lacking a real business model to justify its high operating costs and environmental impact.

The indifference to the approach focused on responsible innovation on the part of US operators, the decision to put on the market solutions that are still unstable and a source of multiple risks, as well as generated in violation of the rules on the protection of the processing of personal data[25] and the protection of intellectual property rights,[26] have led the European legislator to a regulatory reaction aimed at finding a balance between protection and the fascination for economic possibilities (supported in particular during the trialogue phase by France in the Mistral AI case), when an approach based on the precautionary principle might have been more appropriate.

The result of these industrial policy tensions has been the development of a set of rules that distinguish between GPAI models and systems using these models. What is concerning is the so-called systemic risk, which is essentially presumed on the basis of the size of these models and with a relative presumption, with a public registry for GPAI models characterised by systemic risk. The focus is precisely this risk, for which model providers will have to demonstrate that they have carried out adequate analysis and management by monitoring their actions according to the accountability model now dominant in European digital society regulation.

Since these models are intended to be included in AI systems by opera-

[25] See Garante per la protezione dei dati personali, Registro dei provvedimenti n. 112, 30.03.2023, web doc. n. 9870832, https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870832; Garante per la protezione dei dati personali, Registro dei provvedimenti n. 114, 11 March 2023, web doc. n. 9874702, https://www.garanteprivacy.it/web/web/guest/home/docweb/-/docweb-display/docweb/9874702; Garante per la protezione dei dati personali, ChatGPT: Garante privacy, notificato a OpenAI l'atto di contestazione per le violazioni alla normativa privacy, press release of 29 January 24, https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9978020.

[26] See United States District Court, Southern District of New York, *The New York Time Company v. Microsoft Corporation, OpenAI, Inc., OpenAI LP, OpenAI GP, LLC, OpenAI, LLC; OpenAI OPCO LLC, OpenAI Global LLC, OAI Corporation, LLC, and OpenAI Holdings, LLC,* 27 December 2023, https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

tors other than the model developers, their creators must disclose the sources used to train them, in order to comply with transparency obligations (knowledge of the sources can be useful, for example, to identify possible biases).

Finally, also with a view to balancing the management of the potential risks of AI and the desired benefits, the various specific provisions in favour of innovation should be read, starting with the broad exemption foreseen for research activities,[27] up to the *ad hoc* rules on *sandboxes*,[28] i.e, controlled experimentation areas (already adopted in the context of the implementation of the GDPR in several countries), and the provisions aimed at allowing AI products to be tested in the real world with implicit consequences in terms of social experimentation, which, for this reason, also involve a process of ethical evaluation.[29]

IV. Conclusions

Several initiatives, around the world and at different levels, focus on regulating AI. The legislators are trying to provide a first response to the challenges posed by the AI revolution.

The proposed solutions represent a compromise between the protection of fundamental rights and the expected benefits of AI. This has led legislators to only partially address the demand for protection of rights and freedoms of individuals and society, so as not to slow down the development of AI, even more so in those contexts where there is no strong AI industry.

Given this commitment, it is crucial to conduct a thorough interpretative analysis of the regulation and provide guidelines for its implementation. The crucial role of the risk-based approach requires both a harmonised approach consistent with risk management theory and the development of a specific methodology for fundamental rights impact. The latter must be based on key criteria and variables consistent with the AI regulation and the European nor-

[27] See Art. 2.6 ("This Regulation does not apply to AI systems or AI models, including their output, specifically developed and put into service for the sole purpose of scientific research and development").

[28] See art. 57 et seq., AIA. Per *sandbox* regolatoria, l'AIA intende, ai sensi dell'art. 3.55, "a controlled framework set up by a competent authority which offers providers or prospective providers of AI systems the possibility to develop, train, validate and test, where appropriate in real-world conditions, an innovative AI system, pursuant to a sandbox plan for a limited time under regulatory supervision".

[29] See art. 60.3, AIA ("The testing of high-risk AI systems in real world conditions under this Article shall be without prejudice to any ethical review that is required by Union or national law ").

mative framework, starting with the Charter of Fundamental Rights of the European Union, and must be properly applied by the competent authorities and cannot be delegated to standardisation bodies.

In this regard, in relation to several critical comments on the AIA, it is worth noting that in draughting the law it is good to be ambitious, but in the *ex-post* evaluation we need to be realistic. We should put the AIA in context and look back to the years when Europe and many academics advocated a purely ethical approach to AI regulation. We should also recall that the original framework of this regulation was primarily conceived as an industrial safety instrument, with the protection of fundamental rights serving as a mere element of a broader conformity assessment

It is also important to take into account the global context, with powerful governmental and business actors advocating guidelines and solutions other than legal obligations and (as in the case of the GDPR) Cassandras have provided a lengthy list of negative consequences for the EU due to the AIA.

It is on this fine line that the AIA was built, within a legislative process that does not facilitate interaction with non-industry voices, marginalises academia (with the exception of pro-industry voices), and engages civil society in a diffuse way.

The AIA is not the best possible law, but it is a first generation law. Even the first data protection laws were a long way from the GDPR. This is normal in technology regulation; the crossover between economic interest, innovation, and protection of rights requires compromises. Interpretations of this AIA will come to clarify and mitigate its limits; more pointed implementation tools will follow -especially in relation to impact assessment models- and, over the years, new generations of AI laws and a higher level of protection will also come.

THE COUNCIL OF EUROPE CONVENTION ON ARTIFICIAL INTELLIGENCE VERSUS THE EU REGULATION: TWO VERY DIFFERENT LEGAL INSTRUMENTS

Jacques Ziller

Professor of Public Law and European Union law, Universities Paris-1 Panthéon Sorbonne and Pavia

While the institutions of the European Union were working on the regulation of Artificial Intelligence, which gave rise to the AI Act, the Council of Europe (hereafter CoE), which brings together all European states except for Belarus and Russia[1], was also working on this issue. It would be wrong to say that the two organisations have worked in parallel: as the text of the *Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law (Projet de Convention-cadre sur l'intelligence artificielle, les droits de l'homme, la démocratie et l'État de droit / Draft Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*[2] hereafter FCAI) itself shows, there has been and continues to be a great deal of interaction between the institutions of the two European organisations[3]. This is the least we can do, given that all 27 EU Member States are also members of the CoE, along with 19 other states. This is why the EU Council adopted on 21 November 2022 a decision authorising the opening of EU negotiations for a Council of

[1] The Vatican City State has the personality of a sovereign body of public international law, distinct from the Holy See (i.e. the head of the Roman Catholic Church), and enjoys universal recognition, but it has a special nature which explains that, in addition to the Organisations in which the Holy See participates as a permanent observer, such as the CoE, the Vatican City State is a member of only a few IOs, such as the Universal Postal Union (UPU), the International Telecommunications Union (ITU), the International Atomic Energy Agency (IAEA) and the World Tourism Organisation (UNWTO). V. https://www.vaticanstate.va/it/stato-governo/note-generali/origini-natura.html

[2] https://rm.coe.int/cai-2023-28-fr-projet-de-convention-cadre/1680ae19a1; https://rm.coe.int/cai-2023-28-draft-framework-convention/1680ade043 The text adopted at this meeting was not made public until mid-April and has been circulating on social media since 19 March https://rm.coe.int/1680afae3d; https://rm.coe.int/1680afae3c It will then be submitted to the Committee of Ministers, which has the final say, so further changes at this stage are not ruled out. An explanatory statement has also been published https://rm.coe.int/1680afae68; https://rm.coe.int/1680afae67

[3] See for example the CoE news of 12 October 2023 "Secretary General Marija Pejčinović Burić met with European Commissioner for Justice Didier Reynders. The meeting focused on the cooperation between the Council of Europe and the European Union and on the ongoing preparations for the Convention on Artificial Intelligence". https://www.coe.int/es/web/portal/-/secretary-general-meets-european-commissioner-for-justice

Europe Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law[4]. Some recitals of the decision are worth quoting.

> "(4) The Union has adopted common rules which will be affected by the elements to be included in the Convention. These elements include, in particular, a comprehensive set of single market rules applicable to products and services for which AI systems may be used, as well as rules of secondary Union law implementing the Charter of Fundamental Rights of the European Union (EU), taking into account that these rights are likely to be adversely affected in certain circumstances by the development and use of certain AI systems". We will see below the particularities arising from the CoE's and the EU's powers of attribution.

Recital (5) states that the scope of application envisaged for the convention and the AI Act proposal "overlap to a large extent with that legislative proposal in its scope, since both instruments aim to lay down rules applicable to the design, development and application of AI systems, provided and used by either public or private entities". Then in recital (6): "The conclusion of the convention may affect existing and foreseeable future common Union rules or alter their scope within the meaning of Article 3(2) of the (TFEU)". It is striking that this recital refers to Art. 3(2) on the values of the Union, according to which "The Union shall offer its citizens an area of freedom, security and justice without internal frontiers, in which the free movement of persons is assured in conjunction with appropriate measures with respect to external border controls, asylum, immigration and the prevention and combating of crime". As we shall see, the Commission proposal refers only to the legal bases relating to the internal market and not to those relating to controls at the internal and external borders of the Union (Article 77(2) TFEU).

It should also be noted that the European Data Protection Supervisor (EDPS) issued a report on the draft CoE, referred to in a footnote to the Council Decision, with a number of recommendations which have often been repeated in the successive versions of the draft[5]. In his general comments, the EDPS notes that the 'market-centred approach is in line with one of the

[4] Council Decision (EU) 2022/2349 of 21 November 2022 authorising the opening of negotiations on behalf of the European Union with a view to a Council of Europe Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32022D2349

[5] Opinion 20/2022 of the European Data Protection Supervisor on the Recommendation for a Council Decision authorising the opening of negotiations on behalf of the European Union with a view to a Council of Europe Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law (English only) https://www.edps.europa.eu/system/files/2022-10/22-10-13_edps-opinion-ai-human-rights-democracy-rule-of-law_en.pdf

main objectives of the AI Act proposal, the single market dimension of the regulation of AI systems. [At the same time, the scope of competences of the Council of Europe is much broader [...]. In this context, the EDPS considers that the Convention represents an important opportunity to complement the proposed AI Law by strengthening the protection of fundamental rights of all persons affected by AI systems. Therefore [...] the EDPS considers that safeguarding the rights of persons and groups of persons subject to the use of AI systems should be more prominent among the general objectives of the negotiation of the Convention'[6]. As we will also see, the FCAI, since the December 2023 version, includes specific provisions for the European Union, which are clearly the result of the Commission's involvement in the negotiations.

The aim of this contribution is to highlight the advantages and disadvantages of a CoE treaty, such as the FCAI, as opposed to an EU regulation, such as the AI Act. It will then only briefly present the content of the FCAI, a text of which Lorenzo Cotino Hueso rightly says that "the Convention puts *the lyric to the prose* of the AI Act. The AI Act establishes the foundations and structures of a safe and trusted AI ecosystem, the convention focuses on its impact on people and democratic society. The AI Act is methodical, detailed and precise, charting a clear path through technical and legal complexity, setting firm standards and concrete obligations for providers and users or implementers of AI systems. In contrast, on the lyrical side, the convention rises to normatively integrate the fundamental values, ethical principles and human rights that should guide the evolution of AI"[7].

I. The content of the draft Council of Europe Framework Convention

Chapter II of the FCAI is devoted to "General Obligations". According to Art. 4, "Each Party shall adopt or maintain measures to ensure that the activities within the lifecycle of Artificial Intelligence systems are consistent with obligations to protect human rights, as enshrined in applicable international law and in its domestic law". This implies not only the adoption of the necessary regulations and legislation (which, for EU Member States, is partly covered by the AI Act, which is a directly applicable instrument), but also the necessary human and budgetary resources, training and information measures.

[6] Points 10 and 11, p. 7.

[7] Cotino Hueso, L. "El Convenio sobre inteligencia artificial, derechos humanos, democracia y Estado de Derecho del Consejo de Europa", *Revista Administración & Ciudadanía*, EGAP, 2024, Vol. 19.

Article 5 specifies that these are “measures to ensure that Artificial Intelligence systems are not used to undermine the integrity, independence and effectiveness of democratic institutions and processes, including the principle of the separation of powers, respect for judicial independence and access to justice” and that “each Party shall adopt or maintain measures that seek to protect its democratic processes in the context of activities within the lifecycle of Artificial Intelligence systems, including individuals’ fair access to and participation in public debate, as well as their ability to freely form opinions”.

The FCAI establishes the obligation to take measures with regard to the “Integrity of democratic processes and respect for the rule of law” (Art. 5) as well as “to respect human dignity and individual autonomy” (Art. 7). As we will see below, the FCAI is moreover part of the CoE’s core mission, which is to protect, primarily through binding legal instruments, human rights and the rule of law in a democratic society, as laid down in the Statute of the Council of Europe and the ECHR[8].

Chapter III is devoted to the “Principles Related to Activities within the Lifecycle of Artificial Intelligence Systems”, which “sets forth general common principles that each Party shall implement in regard to Artificial Intelligence systems in a manner appropriate to its domestic legal system and the other obligations of this Convention” (Art. 6). These are “human dignity and individual autonomy” (Art. 7), “transparency and oversight” (Art. 8), “accountability and responsibility” (Art. 9), “equality and non-discrimination” (Art. 10), “respect for privacy and personal data protection” (Art. 11), “reliability”, i.e. “measures to promote the reliability of Artificial Intelligence systems and trust in their outputs, which could include requirements related to adequate quality and security throughout the lifecycle of Artificial Intelligence systems” (art. 12), and “safe innovation […] each Party is called upon to enable, as appropriate, the establishment of controlled environments for developing, experimenting and testing Artificial Intelligence systems under the supervision of its competent authorities” (art. 13). As Cotino rightly says, “The Convention not only has a symbolic and meta-legal value, but is also a normative instrument, with the capacity for almost constitutional integration into the legal systems of the States Parties, and has great interpretative potential. This is why the AI Convention surpasses dozens of declarative and soft law instruments that were already superfluous, innocuous and even tedious”.

Chapter IV is devoted to “remedies” and “procedural safeguards”. These

[8] V. Ziller, J. *L’État de droit, une perspective de droit comparé – Conseil de l’Europe*, Brussels, European Parliament Research Service PE 745.673 -2023. https://www.europarl.europa.eu/RegData/etudes/STUD/2023/745676/EPRS_STU(2023)745676_FR.pdf

are obligations of States Parties, not a system of remedies at the CoE level, as we will see below. Chapter V deals with "assessment and mitigation of risks and adverse impacts".

Chapter VI is dedicated to the "implementation of the Convention", with recurrent provisions in recent CoE instruments, relating to non-discrimination (Art. 17), rights of persons with disabilities and of children (Art. 18), public consultation (Art. 19), safeguard for existing human rights (Art. 21), relationship with other legal instruments and wider protection (Art. 22 and 23). Art. 20 "Digital literacy and skills" is more specific to AI: "Each Party shall encourage and promote adequate digital literacy and digital skills for all segments of the population, including specific expert skills for those responsible for the identification, assessment, prevention and mitigation of risks posed by Artificial Intelligence systems".

Chapter VII establishes a "follow-up mechanism and co-operation". As for Chapter VIII on the "final clauses", it is significant that only five States are required to ratify, of which at least three must be members of the CoE, showing the intention to activate the AI Convention as soon as possible.

As Cotino rightly says "Although in general the AI Convention is not characterised by establishing clear obligations and specific rights, there are several reasons to take it normatively into account. [...] I consider the regulation in the AI Convention of general "principles" applicable to all AI systems to be relevant. In this regard, it is worth recalling that, for years, among dozens of declarations and documents, some essential ethical principles of AI have been made visible and distilled. Harvard analysed more than thirty of the main international and corporate declarations on AI ethics and synthesised them into privacy, accountability, security, transparency and explainability, fairness and non-discrimination, human control, professional responsibility, human values and sustainability. The future Convention is positive in that it goes beyond declarations in the realm of *soft law* and regulates these principles, if I may say, it moves from the muses of ethics to the theatre of law. [...] However, there are some concrete elements of the Convention that may go somewhat beyond the AI Act and EU law.

Finally, it is necessary to introduce the specific provisions resulting from the EU Commission's involvement in the negotiations, which we will discuss in section 5, on the particularities of signing and ratifying CoE treaties as opposed to adopting an EU regulation or directive. There are two articles; art. 27 declares that, "1. If two or more Parties have already concluded an agreement or treaty on the matters dealt with in this Convention, or have otherwise established relations on such matters, they shall also be entitled to apply that agreement or treaty or to regulate those relations accordingly, so long as they do so in a manner which is not inconsistent with the object and purpose of this Convention" and "2. Parties which are members of the European Union shall, in their mutual relations, apply European Union rules governing the matters within the scope of this Convention without prejudice to the object and purpose of this Convention and without prejudice to its full application with other Parties. The same applies to other Parties to the extent that they are bound by such rules".

The version proposed before the last Artificial Intelligence Committee meeting in March 2014 also contained a specific provision for the EU in Article 29 -Dispute settlement "If a dispute arises between Parties concerning the interpretation or application of this Convention which cannot be resolved by the Conference of the Parties in accordance with paragraph 1 e of Article 24, the Parties shall seek a settlement of the dispute through negotiation or any other peaceful means of their own choice. The European Union and its Member States shall not, in their mutual relations, avail themselves of Article 29 of the Convention. Nor may the Member States of the European Union invoke this Article of the Convention in any dispute between them concerning the interpretation or application of European Union law". The last two sentences were not included in the version adopted on 14 March; they were in fact a reminder of well-known principles of EU law. As we shall see, these provisions should be read in the light of a possible provision on reservations to the Convention.

II. The reasons for a Council of Europe Treaty on Artificial Intelligence

Remember that the CoE was founded by the Treaty of London of 5 May 1949, signed by ten European states[9] and entered into force on August the 3rd 1949. It is the oldest of the organisations created after the Second World War with the aim of bringing together European countries that share the values of liberal democracy. According to Article 1 of its Statute, the aim of the CoE is to achieve "greater unity among its members in order to safeguard and realise the ideals and principles which constitute their common heritage", including the "primacy of law" (*préminence du droit / rule of law*), and "to facilitate their economic and social progress"[10]. One of the primary objectives of the CoE is the protection of human rights, which led its organs to prepare the

[9] Belgium, Denmark, France, Ireland, Italy, Luxembourg, the Netherlands, Norway, Sweden, the United Kingdom and the United States.

[10] Official translation in the Instrument of Ratification of the Convention for the Protection of Human Rights and Fundamental Freedoms, done at Rome on 4 November 1950,

Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR), which was signed on 4 November 1950 and entered into force on November the 3rd 1953 after being ratified by eight member states; for Spain it was November 24th 1977. The Russian Federation ceased to be a member of the CoE on 16 March 2022, following Russia's military aggression against Ukraine. Belarus is not a full member of the CoE, as it has not signed the European Convention on Human Rights. Its participation in CoE working groups has also been suspended, according to the CoE. On March 17th 2022, the CoE suspended relations with Belarus due to the country's "active participation" in the Russian invasion of Ukraine. Neither Russia nor Belarus has been represented in the *Ad Hoc* Committee on Artificial Intelligence (CAHAI) established on 11 September 2019[11] and, of course, even less in its successor since January 2022, the Committee on Artificial Intelligence (CAI)[12].

The mandate of the Artificial Intelligence Convention was granted from 1st January 2022 until 31st December 2024 with the "CAI Mandate"[13] in the framework of the programme "Effective implementation of the ECHR"[14], which explains the FCAI focus on rule of law and human rights. The mandate was adopted under the authority of the Committee of Ministers, in which CoE member states are generally represented by their Permanent Representative in Strasbourg, exceptionally by their Foreign Ministers. The Committee of Ministers may adopt resolutions and, in particular, recommendations to the governments of the Member States, notably on the follow-up to be given to judgments of the European Court of Human Rights, which are binding on States (ECHR Art. 46). The CoE Statute specifies (Art. 20) the voting procedures. They range from a majority of representatives with unanimity of the votes cast for the most important questions, to a majority of representatives with a two-thirds majority of the votes cast for most resolutions, to a simple majority for questions relating to the Rules of Procedure or the financial and administrative rules.

as amended by Additional Protocols Nos. 3 and 5 of 6 May 1963 and 20 January 1966, respectively, https://www.boe.es/buscar/doc.php?id=BOE-A-1979-24010.

[11] Decision of the Committee of Ministers of the Council of Europe CM/Del/Dec(2019)1353/1.5, 11 September 2019.

[12] https://www.coe.int/fr/web/artificial-intelligence/cai / https://www.coe.int/en/web/artificial-intelligence/cai

[13] https://rm.coe.int/mandat-du-comite-sur-l-intelligence-artificielle-cai-/1680addf7e / https://rm.coe.int/terms-of-reference-of-the-committee-on-artificial-intelligence-for-202/1680a74d2f

[14] https://www.coe.int/fr/web/civil-society/effective-echr-implementation / https://www.coe.int/en/web/civil-society/effective-echr-implementation

The Committee has "tasked the CAI to take into account the key findings and relevant challenges set out in the Secret*ary Ge*neral's 2023 report on the state of democracy, human rights and the rule of law, entitled "Call for a new commitment to CoE val*ues an*d norms"'. The aim was to "establish an international negotiation process and undertake work to finalise an appropriate legal framework on the developmen*t, des*ign, use and decommissioning of Artificial Intelligence, which is based on CoE norms on human rights, democracy and the rule of law, as well as other relevant international standards, and which is conducive to innovation, which may consist of a cross-cutting binding legal instrument including, inter alia, common general principles, as well as additional binding or non-binding instruments to address challenges related to the application of Artificial Intelligence in specific sectors, in accordance with the relevant decisions of the Committee of Ministers". It was also to "maintain a cross-cutting approach by also coordinating its work with other CoE committees and intergovernmental entities that also deal with the implications of Artificial Intelligence in their respective areas of activity, providing guidance to these committees and entities in line with the developing legal framework and assisting them in problem solving", as well as "basing the work on sound evidence and an inclusive consultation process, including with international and supranational partners to ensure a holistic view of the issue". Finally, the aim was to "contribute" to the achievement of the UN 2030 Agenda for Sustainable Development and to review progress in this regard, in particular in relation to Goal 5: Gender equality, Goal 16: Peace, justice and effective institutions".

As summarised by Cotino, "this mandate had a clear intention to transcend borders, seeking to create an "instrument attractive not only to the states of Europe but to the largest possible number of states from all regions of the world", involving "Observers" such as Israel, Canada, the United States, Japan, the Global Partnership on Artificial Intelligence (GPAI), Internet companies, and civil society organisations".

According to Article 30(1) of the FCAI, "This Convention shall be open for signature by the member States of the Council of Europe, the non-member States which have participated in its elaboration and the European Union". Remember that the European Union is party to several CoE treaties, such as the Istanbul Convention on preventing and combating violence against women and domestic violence[15] as of 1st January 2023, and that the accession of the EU to the ECHR is provided for in Article 17 of Protocol

[15] Council of Europe Convention on preventing and combating violence against women and domestic violence (CETS No 210), https://rm.coe.int/1680462543

14 to the ECHR, amending the monitoring system of the Convention and Article 16 TFEU.

Therefore, according to article 31 FCAI -Accesion: "1.After the entry into force of this Convention, the Committee of Ministers of the Council of Europe may, after consulting the Parties to this Convention and obtaining their unanimous consent, invite any non-member State of the Council of Europe which has not participated in the elaboration of this Convention to accede to this Convention by a decision taken by the majority provided for in Article 20.d of the Statute of the Council of Europe, and by unanimous vote of the representatives of the Parties entitled to sit on the Committee of Ministers". There are several CoE treaties to which non-European states have acceded, for example. Canada, Chile, Costa Rica, the Holy See, Japan, Mexico, the United States and the United States are often invited. As for the CoE Convention on Laundering, Search, Seizure and Confiscation of the Proceeds from Crime and on the Financing of Terrorism (CETS No. 198), Morocco, which has also been invited, is the only non-member state to have ratified it. Although it ceased to be a member of the CoE in 2022, the Russian Federation remains a party to several conventions, which it has not denounced, unlike the ECHR.

In our view, the two main reasons for drafting a CoE treaty on Artificial Intelligence were to have a common text for all European states, including the UK after Brexit, and to participate in the global race to be the first to adopt a regulation on Artificial Intelligence, in the hope of serving as a model at least for pluralistic democracies. For example, one can read on the CoE's Artificial Intelligence news page: "On 5-6 March 2024, the Artificial Intelligence Unit of the Council of Europe participated in the OECD-African Union (AU) Dialogue on Artificial Intelligence (AI), sponsored by the UK government and held at the OECD headquarters in Paris, to present the work of the Artificial Intelligence Committee (CAI). The event brought together members of the AU Commission (Algeria, Cameroon, Republic of Congo, Djibouti, Egypt, Ethiopia, Kenya), the AU Working Group on AI and invited experts, including other international organisations with complementary mandates on AI, to discuss the AU Continental Strategy on Artificial Intelligence, AI governance, fostering collaboration and addressing common challenges. Ms Louise Riondel, Co-Secretary of the CAI, participated in the session entitled "The international perspective: from global initiatives to global governance", during which she presented the activities of the Council of Europe in the field of AI, and more specifically the work of the CAI on the Framework Convention on AI and the methodology for assessing the risks and impacts of

AI systems (HUDERIA)"[16]. This is just one of the many activities of the Council of Europe in the field of Artificial Intelligence since 2019, which you can easily find on the website www.coe.int/ai.

An additional reason was obviously to try to propose a text that could be adopted by a large number of other States, including the United States of America. The participation of the latter in the negotiations, in particular for the last meeting of the CAI from 11th to 14th March 2024, had however the effect of reducing its scope of application, in particular because it was finally decided that the Framework Convention would not apply to the private sector. Of course, this in no way prevents a State party to the future Convention from adopting more inclusive legislation, as will be the case for EU Member States through the AI Act.

III. The instrument of the Framework Convention versus the instrument of the Regulation

As mentioned above, the instrument used by the CoE to regulate Artificial Intelligence is a framework convention, i.e. an international treaty known as the *Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*, while the EU uses a regulation known as AI Act[17].

It should be highligted that the term "Ley de Inteligencia Artificial" (Artificial Intelligence Law), used in brackets in the title of the AI Act in the Spanish version of the draft published by the European Commission on 21 April 2021 -and in the version adopted by the European Parliament on 13 March 2024- is incorrect from a legal point of view. The reason is that "European law" does not exist as an instrument of Union law, as the new categorisation of Union acts contained in the Constitutional Treaty of 24 October 2004, which, as is well known, did not enter into force because it had not been ratified by all the Member States, has not been incorporated into the Treaty of Lisbon. The fact that the German, Italian and Dutch (for example) versions of the text also used the word "law" (*Gesetz*, *legge*, *wet*) did not justify the Spanish version established by the European Commission. The Portuguese

[16] https://rm.coe.int/cai-bu-2022-03-outline-of-huderia-risk-and-impact-assessment-methodolo/1680a81e14

[17] Regulation (EU) 2024/... of the European Parliament and of the Council of ... laying down harmonised rules in the field of Artificial Intelligence and amending Regulations (EC) No 300/2008, (EU) Nos.No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Regulation) Official Journal of the European Union, ...

version simply read R*egulamento inteligência artificial*; the French version *législation sur l'intelligence artificielle* is also more correct because it is a legislative act (i.e. adopted by a legislative procedure); the English version *Artificial Intelligence Act* is also correct, as the regulation is a legal act of the Union within the meaning of Article 288 TFEU; likewise the Danish version uses the word R*etsakten*, which means "legal act".

In recent years, it seems that the Commission services tend not to be fussy about the titles of secondary legislation and adopt more of a marketing attitude, using terms that address a non-legal audience; it is also true that the use of English as the main language in institutional *praxis* -although the 24 official and working languages mentioned in Article 55 TEU and Regulation 1/58[18] have the same legal value[19]- allows some ambiguity to be maintained, given that in the UK a law is called an *Act of Parliament*. In fact, the English word that best corresponds to the Spanish word "ley" is *statute*. It should be noted that, unlike the proposed AI Act or the proposed "European *Media Freedom Act*"[20] (*European Media Freedom Act*) in the so-called *Digital Markets Act*[21] *Digital Services Act*[22], it is referred to in brackets as Digital Markets / Digital Services Regulation in most languages, except in German, where the term *Gesetz* is used.

Fortunately, the final text adopted by the Council on 14 May 2024 has been corrected and now reads "Artificial Intelligence Regulation" instead of "Law".

CoE law is simpler from this formal point of view: there is no legal difference between a Convention, a Framework Convention, an Agreement, a Pro-

[18] Regulation No. 1 determining the languages to be used in the European Economic Community, https://eur-lex.europa.eu/legal-content/es/ALL/?uri=CELEX%3A31958R0001

[19] See among others Ziller, J. «Le multilinguisme, caractère fondamental du droit de l'Union européenne», Condinanzi, Canizzarro, Adam et al. (eds.), *Liber amicorum Antonio Tizzano. De la Cour CECA à la Cour de l'Union: le long parcours de la justice européenne*, Torino, Giappichelli, 2018, pp. 1067-1082.

[20] Proposal for a Regulation of the European Parliament and of the Council establishing a common framework for media services in the internal market (European Freedom of the Media Act) and amending Directive 2010/13/EU, https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A52022PC0457

[21] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Regulation) https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A32022R1925

[22] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC (the Digital Services Regulation) https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A32022R2065

tocol, an Arrangement or even a Charter, such as the European Social Charter, until the CoE Statute[23]; they are all public international law treaties. Out of a total of 226 agreements signed at the beginning of March 2024[24] there are three framework conventions: the European Outline *Convention* on *Transfrontier Co-operation between Territorial Communities or Authorities* of 21/05/1980 (ETS n° 106 *Convention-cadre européenne sur la coopération transfrontalière des collectivités ou autorités territoriales / European Outline Convention on Transfrontier Co-operation between Territorial Communities or Authorities*), the Framework *Convention for* the Protection of National *Minorities* of 01/02/1995 (ETS n° 157 *Convention-cadre pour la protection des minorités nationales / Framework Convention for the Protection of National Minorities*) and the Framework Convention on the Value *of* Cultural Heritage for Society of 27/10/2005 (STCE n° 199 *Convention-cadre sur la valeur du patrimoine culturel pour la société / Framework Convention on the Value of Cultural Heritage for Society*). According to the CoE's Directorate-General for Democracy and Human Dignity, "the only difference between "conventions" and "agreements" is the way in which the state can express its consent to be bound. Agreements can be signed with or without reservations as to ratification, whereas conventions, in principle, must always be ratified by the state"[25]. In reality, this indication is also not accurate, since it is the domestic law of each state that determines whether ratification is necessary or whether a simple signature is possible in order to bind the state in question. According to the Explanatory Report of the *Framework Convention on the Value of Cultural Heritage for Society*, for example, "it is a framework Convention that sets out principles and broad fields of action agreed upon by States Parties". However, there are other CoE treaties that state the same. Similarly, the fact that non-members of the CoE can join is not a specific feature of framework conventions. Indeed, reference should be made to recital 11 of the FCAI, according to which "the Convention is intended as a framework and may be complemented by other instruments designed to address specific issues related to the design, development, use and decommissioning of Artificial Intelligence systems".

As we know, the differences between a CoE Treaty and an EU Regulation are essentially due to the fact that the former is binding only on the states

[23] Unless otherwise indicated, translations of Council of Europe texts have been made by the author from the French version.

[24] https://www.coe.int/fr/web/conventions/full-list / https://www.coe.int/en/web/conventions/full-list

[25] https://www.coe.int/fr/web/democracy-and-human-dignity/treaties / https://www.coe.int/en/web/democracy-and-human-dignity/treaties

that have signed and, where appropriate, ratified it, while the latter is binding on all EU Member States -unless there is an exceptional exemption, usually temporary, or based on the protocols relating to Ireland and Denmark (and the UK before Brexit). In addition, there are major differences due to the fact that the EU's competences are much more precisely defined and therefore more limited than those of the CoE.

IV. The limits derived from the respective competences of the Council of Europe and the European Union

In order to avoid errors in the comparison between CoE and EU texts, legal experts are best suited to explain the origin of certain formulations. There is a first essential difference between CoE and the EU, namely, the way in which the competences of the two organisations are formulated and framed.

International organisations do not possess a general power to act; unlike a sovereign state -whose competences are limited only by its obligations under international agreements- the competences of an IO are limited to those conferred by its member states, in accordance with the principle of conferral that applies to intergovernmental organisations -and which has been explicitly mentioned since the Lisbon Treaty in the EU treaties-.

According to Art. 1 of the CoE Statute "a) The purpose of the Council of Europe is to bring about a closer union among its members in order to safeguard and promote the ideals and principles which constitute their common heritage and to further their economic and social progress. b) This purpose shall be pursued through the bodies of the Council by examining matters of common interest, concluding agreements and taking joint action in the economic, social, cultural, scientific, legal and administrative fields, and by safeguarding and enhancing the effectiveness of human rights and fundamental freedoms. c) The participation of Members in the work of the Council of Europe must not alter their contribution to the work of the United Nations and of the other international organisations or unions of which they are members. d) Matters relating to national defence do not fall within the competence of the Council of Europe". In short, the only material limit to the CoE's competences is the exclusion of matters relating to national defence.

As far as the European Union is concerned, a number of provisions should be taken into account: Art. 5(2) TEU which states that "In accordance with the principle of conferral, the Union shall act within the limits of the competences conferred upon it by the Member States in the Treaties to attain the objectives they have defined. Any competence not conferred on the Union in the Treaties lies with the Member States". TFEU which states that "The scope of and arrangements for exercising the Union's competences shall be determined by the provisions of the Treaties relating to each area". The latter wording, introduced by the Treaty of Lisbon, merely puts in black and white what was already clear in the Treaty establishing the European Coal and Steel Community of 1951 and in the Treaties of Rome of 1957 by the precision of their provisions.

When an initiative for EU action is envisaged, the first task of the legal experts in the Commission, the Council and the European Parliament is therefore to check whether there is a legal basis for such action in the treaties. If not, there is a high risk that the acts adopted will be challenged and, sooner or later, annulled by the Court of Justice. A legal basis consists of one or more provisions of the treaties which have the following elements.

Firstly, the action envisaged must fall within an area for which competence has been conferred on the Union. For example, the internal market (Articles 26 and 27 TFEU, as well as 114 and 115, among others), monetary policy (Articles 127 et seq. TFEU), environmental policy (Articles 191 et seq. TFEU), etc. In some cases, competence is conferred implicitly and can be deduced by combining different elements of the "treaty system" as the Court of Justice often uses the expression.

Secondly, action can only be taken to achieve the Union's objectives. These are sometimes specifically mentioned together with the provision referring to the scope of action (e.g. Article 191 TFEU for monetary policy); otherwise they are derived from the more general objectives of Article 3 TEU. Usually, the objectives are set out in carefully chosen wording that sets limits to the policy choices that can be made in the exercise of the competences conferred by the Member States on the Union. When reviewing the legality of secondary law, the Court of Justice checks whether its provisions are consistent with the objectives set out in the Treaties and, if not, annuls the act in question.

Thirdly, only the type of act specified in the relevant provision may be acted upon. The articles of the Treaties often specify whether directives, or regulations, or decisions are to be used, or leave the choice between different acts; alternatively, in many cases they leave a wider choice with the use of the word "measures" (see, for example, for the internal market, Articles 114 -measures- and 115 TFEU -directives-). In any case, even when the word "measures" is used, they can only take the form of acts provided for in the Treaties, as is clear from Art. 288 TFEU.

Fourthly, in order to constitute a legal basis, the relevant provisions must specify the procedure to be followed by the institutions. For the adoption of legislative acts, reference is made to the ordinary legislative procedure, the details of which are specified in Art. 294 TFEU, or a special legislative procedure is explicitly indicated (see, for example, Arts. 114 and 115 TFEU). For non-legislative acts, the procedure to be followed is specified in each case in the relevant Treaty provision (see e.g. Art. 108 and 109 TFEU for State aid control). If the relevant legal bases for the envisaged action do not provide for a type of act that the institutions would wish to use -for example, a regulation instead of a directive- Article 352 TFEU allows such an act to be adopted by a specific procedure requiring a unanimous decision of the Council and the approval of the EP; by contrast, Article 352 cannot be used for action in an area not attributed to the EU. In addition, secondary legislation provides the legal basis for subsequent implementing acts to be adopted by the institutions, bodies, offices and agencies of the Union and, where appropriate, by the authorities of the Member States. These implementing acts must comply with the provisions of the relevant secondary legislation and, in the first instance, with the objectives set out in the body of the Union act or in its introductory recitals.

It is essential to take the above into account in order to understand the framework of EU law applicable to Artificial Intelligence. Indeed, given the large number of proposed Commission acts and communications relating to digitalisation and AI published in recent years, there is a risk of forgetting the limits that the principle of conferral imposes on the EU institutions.

A typical example is the "Ethical guidelines for the use of Artificial Intelligence (AI) and data in education and training for educators" published by the Commission on 25 October 2022[26]. Reading this document, as well as the description of the so-called "European Education Area"[27], it seems as if the European Commission acts somewhat like a European Ministry of Education and Universities. Among other things, it explains that "The idea of creating a European Education Area was first endorsed by European leaders at the 2017 Social Summit in Gothenburg, Sweden. The first packages of measures were adopted in 2018 and 2019. [In September 2020, the Commission set out in a Communication its renewed vision for the European Education Area and concrete measures to achieve it. The Council of the EU responded with the Resolution of February 2021 on a strategic framework for European cooperation in education and training for the period 2021-2030". Those unfamiliar with EU legislation might expect European legislation relating precisely to education.

[26] https://op.europa.eu/es/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1

[27] https://education.ec.europa.eu/es/about-eea/the-eea-explained

However, Article 165 TFEU, the only possible legal basis for such action, specifies that the ordinary legislative procedure shall be used for the adoption of "incentive measures, excluding any harmonisation of the laws and regulations of the Member States", which drastically reduces the Union's competences in this area. It is true that Member States remain free to give a certain legal scope to the so-called *soft law* documents adopted by the institutions. However, such a reference does not mean that an instrument of EU law applies.

The text of the CoE Statute is very simple in its application, compared to the acrobatics involved in finding an adequate legal basis in EU law and ensuring that the text does not go beyond what the principle of attribution allows.

As underlined in the explanatory memorandum of the Commission's proposal "the legal basis for the proposal is, first of all, Article 114 of the Treaty on the Functioning of the European Union (TFEU), which deals with the adoption of measures to ensure the establishment and functioning of the internal market. This proposal is a key part of the EU's Digital Single Market Strategy. Its primary objective is to ensure the proper functioning of the internal market by establishing harmonised rules, in particular as regards the development, placing on the Union market and use of products and services using AI technologies or delivered as stand-alone AI systems. Some Member States are already considering national rules aimed at ensuring that AI is safe and is developed and used in accordance with fundamental rights obligations. This is likely to cause two fundamental problems: (i) fragmentation of the internal market as regards essential elements, in particular the requirements applicable to AI products and services, their marketing, their use, and the responsibility and oversight of public authorities; and (ii) a significant decrease in legal certainty for providers and users of AI systems as to how existing and new rules in the Union will apply to such systems. Given the extensive cross-border movement of products and services, these two problems can best be addressed through EU harmonisation legislation".

To be more precise, this is Article 114(1) TFEU: "The European Parliament and the Council shall, acting in accordance with the ordinary legislative procedure and after consulting the Economic and Social Committee, adopt the measures for the approximation of the provisions laid down by law, regulation or administrative action in Member States which have as their object the establishment and functioning of the internal market", then of the achievement of the objectives set out in Article 26(2) TFEU on the internal market: "The internal market shall comprise an area without internal frontiers in which the free movement of goods, persons, services and capital is ensured in accordance with the provisions of the Treaties."

The Explanatory Memorandum adds that 'in addition, given that this proposal contains certain specific rules for the protection of individuals with regard to the processing of personal data, notably restrictions on the use of AI systems for "real-time" remote biometric identification in publicly accessible areas for law enforcement purposes, it is appropriate to base this Regulation, as far as these specific rules are concerned, on Article 16 TFEU: "Everyone has the right to the protection of personal data concerning them." "The European Parliament and the Council, acting in accordance with the ordinary legislative procedure, shall lay down the rules relating to the protection of individuals with regard to the processing of personal data by Union institutions, bodies, offices and agencies, and by the Member States when carrying out activities which fall within the scope of Union law, and the rules relating to the free movement of such data. Compliance with these rules shall be subject to the control of independent authorities."

It is striking that among the documents cited in the appendix to the Explanatory Memorandum there is an Annex IX *Union legislation on large-scale IT systems in the area of freedom, security and justice*. For example, it refers to the Regulation of 30 November 2017 establishing an Entry-Exit System (EES)[28]. The legal bases for this Regulation are "the Treaty on the Functioning of the European Union, and in particular Article 77(2)(b) and (d) thereof, concerning 'the checks to which persons crossing the external borders shall be subject' and 'any measure necessary for the gradual establishment of an integrated management system for external borders'; and Article 87(2)(a) thereof, concerning 'the collection, storage, processing, analysis and exchange of relevant information' in the field of police cooperation. Since these legal bases also provide for recourse to the ordinary legislative procedure, it is questionable why these provisions are not also cited in the text proposed by the Commission. Moreover, Article 87(3) TFEU provides that for cooperation in operations between operational cooperation between "police, customs and other specialised law enforcement services in relation to the prevention, detection and investigation of criminal offences." The Council shall act unanimously after consulting the European Parliament. And Art. 77(3) provides that "the Council, acting in accordance with a special legislative procedure, may adopt

[28] Regulation (EU) 2017/2226 of the European Parliament and of the Council of 30 November 2017 establishing an Entry-Exit System (EES) for recording entry and exit data and refusal of entry data of third-country nationals crossing the external borders of the Member States, determining the conditions of access to the EES for law enforcement purposes and amending the Convention implementing the Schengen Agreement and Regulations (EC) No 767/2008 and (EU) No 1077/2011. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32017R2226

provisions concerning passports, identity cards, residence permits or any other such document. The Council shall act unanimously after consulting the European Parliament."

Perhaps this is why the Commission has avoided citing these two articles, which apply, among other things, to the system of governance of AI Act provisions.

To all those who criticise the Commission's draft for not being broad enough on AI and for giving too much weight to data protection, it suffices to remind them of this point. In particular, Article 114 requires a link to be found with the internal market, i.e. the four freedoms of movement, and does not allow the adoption of a text binding on the EU institutions, but only allows the adoption of a text binding on the Member States. In contrast, Art. 16 does. This explains why, unlike the GDPR[29] which is based on Art. 16 TFEU, the 2001 Regulation on public access to documents[30] applies only to the institutions, bodies, offices and agencies of the Union and not to the Member States. The latter is based on Art. 15 TFEU (ex Art. 255 TEC) which states, inter alia, that "general principles and limits on grounds of public or private interest governing this right of access to documents shall be determined by the European Parliament and the Council, by means of regulations, acting in accordance with the ordinary legislative procedure".

Therefore, it is not surprising that the draft's recitals and provisions, which directly affect public authorities, are highly complex.

V. The need to ratify the Council of Europe Treaty versus the direct applicability of the EU regulation

Unlike EU directives, regulations and decisions of a general nature, which in principle apply directly to all member States, CoE treaties only apply to states that have signed and ratified them if their constitution so requires.

The ECHR is the CoE's most important instrument in general terms, starting with the rule of law, fundamental rights and freedoms and democ-

[29] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation) https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32016R0679

[30] Regulation (EC) No 1049/2001 of the European Parliament and of the Council of 30 May 2001 regarding public access to European Parliament, Council and Commission documents https://eur-lex.europa.eu/legal-content/ES/ALL/?uri=celex%3A32001R1049

racy. Unlike its other conventions and agreements -including the additional protocols to the ECHR- accession to the ECHR is binding on all CoE member states, making it binding on all forty-six CoE states and thus on all EU member States. On the other hand, the FCAI, as is usually the case with CoE treaties, will only be binding on those states that have ratified it, as it follows from Art. 30(3) that only five states are required to ratify, of which at least three must be CoE members.

Unless otherwise specified, States Parties may make reservations or declarations to CRC Conventions at the time of signature or when depositing the act of ratification. The object and effect of a reservation or declaration may be to specify how a treaty is to be applied in relation to a State Party. Reservations of a general nature are not permitted in respect of the ECHR; a reservation may only be made in respect of a particular provision of the Convention "to the extent that a law for the time being in force in its territory is not in conformity with that provision". The draft FCAI provides in Art. 34 -Reservations "No other reservation may be made in respect of this Convention" with a single exception provided for in Art. 33 concerning federal States, which might be necessary for states that are not members of the CoE, such as the United States of America or Canada in particular. It should be noted that the provisions of Art. 27 -Effects of the Convention on EU Member States make it possible to avoid reservations on their part or on the European Union.

The ECHR is directly applicable in most States Parties. Direct application means that the Convention can be invoked before all national courts. This does not mean that the institutions of the State concerned -legislature, administration and judiciary- are not bound to respect the Convention, but that it may be more complicated for an individual to enforce the rights guaranteed by the Convention. The case of FCAI is more delicate. On the one hand, it will have to be determined to what extent its provisions are sufficiently precise to be considered *self-executing*, which will vary from one State Party to another. In particular, there are States where the courts themselves rule on the interpretation of a treaty, and others where they request an interpretation from the Ministry of Foreign Affairs. In addition, there are States where the constitution explicitly considers treaties to be superior to the law, and others where the question is not settled. This will lead to a non-uniform application of the provisions of the FCAI, especially as, unlike the ECHR or the Charter of Social Rights, the FCAI does not provide for a judicial body such as the European Court of Human Rights or a quasi-judicial body such as the European Committee of Social Rights[31] to settle disputes arising from its application.

[31] Salcedo Beltrán, V. C. "La Carta Social Europea y el procedimiento de reclamaciones

That said, both the Court and the Committee refer in their jurisprudence to all relevant instruments of the CoE, as well as to other instruments of international law as useful context for the exercise of their jurisdiction.

VI. In conclusion

As a conclusion on the possible impact of the CoE binding Act on AI, it should be added that the case law of the European Court of Human Rights is set to evolve on the subject of Artificial Intelligence. A first sign of this is the dissenting opinion of Judge Darian Pavli on the judgment of 4 June 2019 in Case 39757/15 *Sigurður Einarsson and Others v. Iceland*[32] concerning the investigation of possible criminal acts linked to the financial crisis and, if appropriate, the prosecution of the persons concerned who are members of the management bodies of one of the largest Icelandic banks, Kaupþing banki.

In the view of the majority of the Chamber, despite frequent complaints to the prosecutor about the lack of access to documents, at no time do the applicants appear to have formally requested a court to grant access to the "full collection of data" or to carry out further investigations, or to have suggested further investigative steps -for example, a new search using keywords suggested by them-. In this regard, the Irish Supreme Court notes the government's submission that the evidence before the Court of First Instance included a general description of the objects seized and their approximate contents. In these circumstances, and given that the applicants did not specify the type of material they sought, the European Court is satisfied that the lack of access to the data did not deprive the parties of a fair trial.

Judge Pavli, in paragraph 21 of his opinion, said: "With all due respect to my colleagues, this argument, in my view, considerably underestimates the complexity of analysing large interconnected amounts of research data, whether "mere" human intelligence or with the assistance of Artificial Intelligence". One swallow does not a summer make, but it is very likely that the Strasbourg Court will increasingly have to rule on issues relating to the use of Artificial Intelligence systems, as it did on data protection, where it relied in particular on Art. 8 of the Right to respect for private and family life, and that it will take due account of the FCAI, as well as the AI Act and national laws and regulations, in constructing its jurisprudence.

colectivas: un nuevo y excepcional escenario en el marco legislativo laboral", *Trabajo y Derecho* 91-92/2022, pp. 1-36.

[32] https://hudoc.echr.coe.int/fre?i=001-193738

ARTIFICIAL INTELLIGENCE REGULATION FROM OUTSIDE THE EUROPEAN UNION: REGULATORY IMPULSES FROM OTHER PARTS OF THE WORLD AND A VIEW FROM IBEROAMERICA

Juan Gustavo Corvalán – María Victoria Carro

Director of the Innovation and Artificial Intelligence Laboratory of the Faculty of Law of the University of Buenos Aires – PhD Candidate, University of Genoa. Research Director, UBA IALAB

I. Introduction

2023 has witnessed momentous changes in the field of Artificial Intelligence (hereafter AI), with all relevant actors doing their part to accelerate a transformation that has had an impact on almost all areas of knowledge and our daily lives.

On the one hand, following the success of ChatGPT, the technological giants have launched themselves into a race not only for multimodal generative AI, but also for systems that are capable of captivating a larger number of users and winning their preference. Then, the big thinkers and researchers in the field, who have been responsible for sounding the alarm and spreading the great risks of this type of technology, have even expressed their concerns in an open letter that sought to halt its development for a period of time that has already expired[1]. Finally, there are the users, who learn to take advantage of the new tools, either as end consumers or by reusing them in creative ways to perfect precision in specific tasks, thus contributing to the flourishing of the ecosystem.

But that is not all. 2023 has also been the year in which AI regulations have become, from a mere future will, a priority on the agenda. At the global level and in the face of the scenario described in the previous paragraph, states have started to think about concrete measures and to seek an active approach, even by those who traditionally decided to take a non-intervention stance[2].

[1] On the open letter expand on: "The letter in which more than 1,000 experts call for curbing Artificial Intelligence as a "threat to humanity"" *BBC News World* (2023), available at: https://www.bbc.com/mundo/noticias-65117146 (accessed 24 January 2024).

[2] For example, the United States has been criticised for its passivity in the face of such technologies. See Knight Will, 'Rain of criticism for countries ignoring AI" *MIT Technology Review*, (2019), available at: https://www.technologyreview.es/s/10939/lluvia-de-criticas-para-los-paises-que-estan-ignorando-la-AI (accessed 24 January 2024).

First, after intense technical and political work, the European Union succeeded in drafting the final text of the world's first Artificial Intelligence Law. It is expected to enter into force in the remainder of the first half of 2024[3].

For its part, the United States was not left out of the wave of regulation and AI finally entered the political debate more forcefully. But it was not just words. The discussion culminated in President Biden's Executive Order on AI at the end of October 2023, which aims to improve AI safety. However, such instruments can be revoked at any time by another president. Unlike the vigorous work being done by EU representatives to reach consensus, the US instrument lacks the legitimacy of Congress, which is in fact so polarised that it is unlikely to produce any meaningful AI legislation anytime soon[4].

China has also emerged as a major player in this regulatory landscape. In fact, it has been one of the most advanced states in recent times, enacting individual laws on different aspects and risks that are gaining importance related to these technologies. Its latest instrument has been in relation to generative AI, which addresses issues such as data privacy and intellectual property. However, in June 2023, China's State Council announced a change of approach: a comprehensive and all-encompassing Artificial Intelligence law similar to that of the EU is on the way[5].

Finally, from Latin America we anticipate that our continent is gradually beginning to participate in a progressive manner in efforts to regulate AI. In general terms, the countries have strategic plans on the one hand and ethical recommendations on the other. In addition, all of them have more or less up-to-date data protection laws. What is new, however, are some general draft laws, such as those of Mexico and Brazil, which we will examine in greater depth in the following lines.

At UBA IALAB, we have recently made a significant effort to segment and analyse these documents in order to have a detailed perspective of what is happening at regional and global level[6]. This research has served as input to

[3] "European Parliament moves ahead with legislation to regulate Artificial Intelligence" *The Observer,* (2024), available at: https://www.elobservador.com.uy/nota/el-parlamento-europeo-avanza-con-la-legislacion-para-regular-la-inteligencia-artificial-2024213104726 (accessed 22 February 2024).

[4] Ryan-Mosley Tate, "US vs Europe: Biden takes the lead in the race to regulate AI" *MIT Technology Review*, (2023).

[5] Ryan-Mosley Tate, "*Around the world for AI regulations in 2024", MIT Technology Review*, (2024), available at: https://www.technologyreview.es/s/16069/vuelta-al-mundo-por-las-regulaciones-de-la-AI-en-2024 (accessed 22 February 2024).

[6] See Corvalán Juan G. (direction), Sánchez Caparrós Mariana, Rabán Melisa (coordination), Stringhini Antonella, Papini Carina Mariel, Heleg Giselle, Bonato Valentín, "Pro-

be presented at the "Conference on regulation and legislation of Artificial Intelligence: generative AI and international tendencies" held on 5 June 2023 in the Chamber of Deputies of Argentina, where a series of recommendations were discussed and developed as a roadmap to address a possible regulation of the use and implementation of Artificial Intelligence in Argentina[7].

Throughout this paper, we will discuss some of the most relevant and problematic aspects of these and other regulatory initiatives, including a special development of trends in some Latin American countries. In this sense, we will seek to broaden our understanding of the effects, both positive and negative, of these regulations in diverse and varied sectors and socio-economic contexts. In short, the purpose will be to highlight the direction in which legislative efforts are heading at the global level.

II. Getting out of the pot in time and other regulatory challenges

How to regulate something that is constantly changing? How to control massive, macro and often imperceptible effects? Europe put an end to these unknowns in an attempt to get out of the pot in time, according to the famous parable of the boiled frog[8], and many other states followed this trend so as not to end up as prisoners of boiled water. In reality, there were many conjunctural factors that contributed to the fact that, at a global level, the regulation of AI went from being a mere wish for the future to a priority on the agenda.

Prior to 2023, states and international organisations limited themselves to issuing "*soft law*" in the form of ethical suggestions or recommendations to relevant actors in the field of AI[9]. These are guiding principles of general

puestas de regulación y recomendaciones de inteligencia artificial en el mundo. Síntesis de principales aspectos" *IALAB UBA*, (2023), available at: https://ialab.com.ar/wp-content/uploads/2023/08/Propuestas-de-regulacion-y-recomendaciones-de-AI-en-el-mundo-1.pdf (accessed 9 March 2024).

[7] On the consensus and recommendations resulting from the conference, see: "Puntos de partida para la regulación de la inteligencia artificial en Argentina" in Corvalán Juan G. (director), "Tratado de Inteligencia Artificial y Derecho" *Thompson Reuters La Ley*, (2023), 2nd edition.

[8] The famous parable of the boiled frog teaches us that if we put one of these amphibians in a pot of boiling water, it immediately tries to get out. On the other hand, if we place it in water at room temperature, and do not frighten it, it remains calm. As the temperature rises, it will stay there and do nothing, but it will become increasingly dazed until the water boils and it is no longer able to escape. While its internal apparatus for detecting threats to survival is prepared for sudden changes in the environment, it is not capable of detecting slow, gradual effects.

[9] Examples of such ethical documents include: the White Paper on Artificial Intelligence

scope, as the imposition of mandatory requirements was often considered excessive and even hasty in a constantly evolving field. There are two reasons for this. First, because "preventive" regulations could hinder innovation and consequently its benefits[10]. Second, because some industries have managed to regulate themselves successfully guided by cultural and institutional pressure[11].

However, in this case, the various pressures were slow in coming. Similar to what happens with personal data, users tend to downplay impacts on the rights that they cannot see or perceive directly. By the end of 2022, it was likely that any average person unfamiliar with the tech industry would conceive of AI as something that was yet to come, that reasonably we could expect in the future. That is without bearing in mind that your Netflix account was already employing AI to help you choose the next series, or that through this technology Instagram filters could recognize your face.

However, over the last year, the rise of large language models (LLMs) -or, in other words, *Foundation Models*- has changed everything. In particular, the arrival of OpenAI's GPT-4 led some experts to believe that we are facing a kind of prelude to superintelligence[12] and, along with it, its respective spectrum of risks.

produced by the European Commission in 2020, the first set of Intergovernmental Policy Guidelines on AI adopted in 2019 by the 36 OECD partner countries, the Ethical Guidelines for Trustworthy AI created by the High Level Expert Group on AI constituted by the European Commission in 2019, and the UNESCO Recommendation on the Ethics of Artificial Intelligence adopted by Member States in 2021.

[10] O'sullivan Andrea, "If governments control Artificial Intelligence too much we will lose its benefits" *MIT Technology Review*, (2017), available at: https://www.technologyreview.es/s/9688/si-los-gobiernos-controlan-demasiado-la-inteligencia-artificial-perderemos-sus-beneficios (accessed 28/2/2024).

[11] AI self-regulation has been a proposal to avoid overly restrictive regulations. See Páez Giménez Efrén, "ExCEO of Google proposes self-regulation in Artificial Intelligence" *DPL News*, (2023), available at: https://dplnews.com/exceo-de-google-propone-autorregulacion-en-inteligencia-artificial/ (accessed 29 February 2024).

Similarly, DeepMind co-founder Mustafa Suleyman has said that while top-down regulation is needed, there are examples of industries that have successfully self-regulated. See Douglas Heaven Will, "DeepMind's cofounder: Generative AI is just a phase. What's next is interactive AI*" MIT Technology Review*, (2023), available at: https://www.technologyreview.com/2023/09/15/1079624/deepmind-inflection-generative-ai-whats-next-mustafa-suleyman/?utm_source=LinkedIn&utm_medium=tr_social&utm_campaign=site_visitor.unpaid.engagement (accessed 29 February 2024).

[12] Romero Sarah, "Microsoft claims GPT-4 can reason like a human" *Muy Interesante*, (2023), available at: https://www.muyinteresante.com/actualidad/60456.html (accessed on 3 March 2024). Also see: Bubeck et. al, "Sparks of Artificial General Intelligence: Early experi-

In addition, users began to notice how some of the most sophisticated intelligent systems were creeping into various tasks in their daily lives. At UBA IALAB we have documented the impact on productivity in multiple tasks, along with other studies that also show how this technology is changing the way we work on a large scale[13].

These seasonings and others, finally led to the creation of a widespread awareness of the importance of controlling AI, which included putting existential risk issues increasingly on the agenda. In this way, the uncertainties about AI regulation that once seemed almost impossible obstacles to overcome became necessary to address and resolve in some way. From peripheral debates that occasionally deserved the occasional opinion piece, they have become the center of public and political discussion.

So far, it is already possible to identify the first tendencie. The AI principles adopted by the OECD in 2019 served throughout this time as a global reference to guide the rest of the ethical recommendations of both international organisations and governments, and now also constitute a starting point that helps these same actors to shape a human-centred regulation and the democratic values for a reliable AI.[14]

In the following, we will present some further orientations that were taken into account and also which others were left out. In doing so, it is important to bear in mind that when asked about the approach to regulation, there is no generally accepted taxonomy, but that the answer depends on different aspects, which are presented here in the form of dichotomies, e.g., unity vs. fragmentation, or "hard law" vs. "soft law". In the middle of each of these, there is a range of nuances that materialise in the decisions actually taken by individual states to direct their regulation.

1. Unity vs. fragmentation: "Horizontal" or "vertical" approach?

One of the big debates in creating a framework for AI standards revolves

ments with GPT-4", *arXiv:2303.12712*, (2023), available at: https://arxiv.org/abs/2303.12712 (accessed on 3 March 2024). Also see Aguera and Arcas Blaise, Norving Peter, "Artificial General Intelligence Is Already Here" *NOEMA*, (2023), available at: https://www.noemamag.com/artificial-general-intelligence-is-already-here/ (accessed 18 March 2024).

[13] See Corvalán Juan G., Díaz Dávila Laura Cecilia, Guilera Soledad, Le Fevre Enzo (address), "La revolución de la productividad. Cómo impacta la IAGen y ChatGPT en la reducción de tiempos y en la optimización de las tareas" *UBA IALAB*, (2024), available at: https://ialab.com.ar/wp-content/uploads/2024/02/Resumen-Ejecutivo.pdf (accessed 18 March 2024).

[14] Morini Bianzino et. al, "The Artificial Intelligence (AI) global regulatory landscape. Policy trends and considerations to build confidence in AI" *EY*, (2024).

around how to regulate something that is so heterogeneous. From facial recognition systems to autonomous cars to predictive and generative systems, the spectrum of tools, functionalities, techniques and levels of autonomy is so varied that experts often question whether a general and comprehensive regulation is the right approach.

First, this dichotomy is also related to the position of those who remind us that rules already exist to regulate AI. To argue that it is necessary to dictate rules that apply to intelligent systems is to ignore large swathes of existing law because, in fact, such regulations already exist, however imperfectly[15]. Liability, contract and intellectual property frameworks are applicable even in cases involving autonomous technologies. However, this does not mean that existing legal arrangements are optimal or that difficulties of interpretation and application do not arise in the face of increasingly complex situations.

Another way of framing this duality is between "horizontal" or "vertical" approaches. In a horizontal perspective, regulators create a comprehensive regulation that covers the many impacts that AI can have. In a vertical strategy, policymakers adopt a tailored approach, creating different regulations to address different applications or types of AI[16]. In this sense it has been said that the EU's AIA is horizontally tilted and China's regulations are vertically tilted[17].

In the introduction to this paper, we noted that China has taken a distinctive approach to regulating AI, enacting individual laws on different aspects and particular risks, such as *deepfakes*. This practical approach has made the Asian giant the best candidate to react quickly and respond to changes brought about by new technologies. So much so that China was probably the first country in the world to introduce legislation on generative AI, just a few months after the major eruption of ChatGPT[18]. However, in June 2023 the

[15] Cuellar Mariano-Florentino, "Reconciling Law, Ethics, and Artificial Intelligence: The Difficult Work Ahead" *Stanford University Human-Centered Artificial Intelligence*, (2019), available at: https://hai.stanford.edu/news/reconciling-law-ethics-and-artificial-intelligence-difficult-work-ahead (accessed 29 February 2024).

[16] O'Shaughnessy Matt, Sheehan Matt, "Lessons From the World's Two Experiments in AI Governance", *Carnegie Endowment for International Peace*, (2023), available at: https://carnegieendowment.org/2023/02/14/lessons-from-world-s-two-experiments-in-ai-governance-pub-89035 (accessed 3 March 2024).

[17] O'Shaughnessy Matt, Sheehan Matt, "Lessons From the World's Two Experiments in AI Governance", *Carnegie Endowment for International Peace*, (2023), available at: https://carnegieendowment.org/2023/02/14/lessons-from-world-s-two-experiments-in-ai-governance-pub-89035 (accessed 3 March 2024).

[18] Yang Zeyi, "Four things to know about China's new AI rules in 2024" *MIT Technology*

State Council announced a change of course: a comprehensive and all-encompassing Artificial Intelligence law is on its way. As expected, it does not seem that the text will come as fast as the aforementioned specific rules we were used to.

The same duality applies to the bodies in charge of oversight and enforcement. Many states are now expressing, through their regulatory initiatives, their willingness to create a centralised and specialised body to ensure effective enforcement of AI rules. Recently, the EU has launched its Office for Artificial Intelligence to promote the use of reliable AI[19] . In its draft law, Mexico plans to establish an entity known as the Mexican Council of Ethics for Artificial Intelligence and Robotics (CMETIAR) to propose laws and monitor compliance[20] .

Even the United States, which has always prided itself on having a patchwork of federal and state authorities examining their parts of these technologies, has considered the idea of concentrating these powers in a single agency. In a hearing before Congress, senators from both parties and Sam Altman, CEO of OpenAI, argued that a new federal agency was needed to protect citizens from harmful AI[21].

The problem with such an initiative at this point is the overlap with the existing work of other public bodies, both at the state and federal level. For example, the National Highway Traffic Safety Administration is in charge of autonomous cars and the Department of Homeland Security has issued reports on potential threats to critical infrastructure from these technologies. Likewise, the Federal Trade Commission and the Food and Drug Administration regulate how companies use AI. In addition, there has been a history of bills that some legislators have abstained from voting on because their enactment would override state legislation on the subject[22]. As if that were

Review, (2024), available at: https://www.technologyreview.com/2024/01/17/1086704/china-ai-regulation-changes-2024/ (accessed 1 March 2024).

[19] "EU launches its Artificial Intelligence Office to promote its reliable use", *El Tiempo*, (2024), available at: https://www.eltiempo.com/tecnosfera/novedades-tecnologia/la-union-europea-inaugura-su-oficina-de-inteligencia-artificial-857260 (accessed 1 March 2024).

[20] González Fernanda, "Presentan propuesta de ley para regular la AI en México" *The Wired*, (2023), available at: https://es.wired.com/articulos/diputado-presenta-propuesta-de-ley-para-regula-la-AI-en-mexico (accessed 1 March 2024).

[21] Johnson Kari, "Scared by ChatGPT, US lawmakers want to create an AI regulatory body", *The Wired*, (2023), available at: https://es.wired.com/articulos/legisladores-de-ee-uu-quieren-crear-organismo-regulador-de-inteligencia-artificial-y-chatgpt (accessed 1 March 2024).

[22] This is what has happened with California lawmakers deciding on federal privacy regulation. See: Johnson Kari, "Scared by ChatGPT, US lawmakers want to create an AI regulatory body", *The Wired*, (2023), available at: https://es.wired.com/articulos/legisla-

not enough, it has been argued that a decentralised or fragmented approach avoids hampering the industry[23].

While some states tend to adopt a more extreme aproximation, in reality, they adopt a dual approach that is both cross-sectoral and sector-specific[24]. The first approach, cross-sectoral, provides a framework of fundamental safeguards, regardless of the sector in which the AI is developed or used. The second, sector-specific approach establishes additional guidelines or obligations for the use of AI to address risks and vulnerabilities within specific domains[25]. While the first framework tends to be short, not very detailed, and aims to be long-lasting, comprehensive regulation comes from experts close to the field of application.

Singapore's Model AI Governance Framework, for example, provides industry-independent guidance to private organisations to align with guiding principles on the ethical use of AI. Complementarily, the Monetary Authority of Singapore (MAS) issued sector-specific guidance to the financial sector on fairness, ethics, accountability and transparency in the use of AI and data analytics[26].

2. Second. Mandatory vs. voluntary and the strengthening of collaboration

To avoid a disappointing regulation, policymakers and legislators who promote and participate in the development of regulations that control AI need to have a thorough understanding of the technology[27]. The risks of AI are often exaggerated and myths propagated, leading to an overestimation of the urgency and stringency needed in regulation. In turn, this leads to disproportionate and misguided responses in the absence of a critical and expert assessment that takes into account, among other aspects, the underlying interests at play.

dores-de-ee-uu-quieren-crear-organismo-regulador-de-inteligencia-artificial-y-chatgpt (accessed 1 March 2024).

[23] O'sullivan Andrea, "If governments control Artificial Intelligence too much we will lose its benefits" *MIT Technology Review*, (2017).

[24] Morini Bianzino et. al, "The Artificial Intelligence (AI) global regulatory landscape. Policy trends and considerations to build confidence in AI" *EY*, (2024).

[25] Morini Bianzino et. al, "The Artificial Intelligence (AI) global regulatory landscape. Policy trends and considerations to build confidence in AI" *EY*, (2024).

[26] Morini Bianzino et. al, "The Artificial Intelligence (AI) global regulatory landscape. Policy trends and considerations to build confidence in AI*" EY,* (2024).

[27] Knight Will, "Politicians need to understand how AI works (urgently)*" The Wired*, (2023), available at: https://es.wired.com/articulos/inteligencia-artificial-los-politicos-deben-aprender-rapido-sobre-AI (accessed 1 March 2024).

In reality, not only a multidisciplinary but also a multi-sectoral dialogue is needed. The real challenge of AI regulation is how to reconcile and collaborate between two very different worlds: on the one hand, the political bodies of states and their respective advisors, with a perspective focused on impact and social needs, who spend time negotiating and seeking consensus in order to ensure safe AI through standards such as transparency. On the other hand, the more or less large companies of the technology industry, immersed in a race for innovation trying to protect their economic interests, more aware than anyone else of the limitations and potential of the technology.

Some very obvious and already theorised tensions arise in relation to the above, such as the difficulties of legislators to understand some basic technical concepts, and the rapid pace at which companies are advancing and which, in turn, governments and their laws are not able to keep up with[28]. However, apart from that, a more challenging issue is the fact that compliance with requirements that laws impose according to the state of the art could in many cases depend purely and exclusively on the will of the technological giants. We will see.

While state action on AI control was -and in many cases is- limited to ethical documents, private companies have little incentive to concern themselves with AI ethics. First, because all the requirements and safeguards proposed by international bodies, such as producing and publishing information, involve investing money. If they then have to make their projects transparent, this investment might be justified. But if, on the other hand, they are not required to provide information even on smart systems developed for the public sector, then, from a business point of view, whether they follow the ethical requirements or not, is simply irrelevant. At most, doing so might make for a good marketing campaign.

It cannot be overlooked that never before have values and business ethics in the face of AI been such a recurring theme within the economic and social sphere as they are today. While small *startups* may fly under the radar, large consultancies and tech giants are leading the way by being constantly under scrutiny, or so it seems. Elon Musk was one of the first to call for regulation. The open letter also echoed this issue[29]. In May 2023, Sam Altman called on

[28] Jonhson Bobbie, "The AI legal problem: How to regulate something that keeps changing" *MIT Technology Review*, (2019).

[29] The open letter to pause AI development stated that the pause must be public and verifiable, and include all key stakeholders. If such a pause cannot be implemented quickly, governments should step in and institute a suspension. See "The letter in which more than 1,000 experts call for a halt to Artificial Intelligence as a threat to humanity" *BBC News World*, (2023), available at: https://www.bbc.com/mundo/noticias-65117146 (accessed 18 March 2024).

the US Congress to regulate AI[30]. Google CEO Sundar Pichai was not far behind and reiterated the call[31].

It is also true that these calls for state action are followed and justified by warnings about certain risks, -if not existential, very serious- that technology could cause in our society. It is curious that the people who are calling for limits and raising alarms are the same people who create these tools. Probably, presenting oneself as the creator of the AI that comes closest to superintelligence to date, can convince users that one's service is more powerful and better than that of competitors[32].

The point is that while ethical efforts are valuable, they will never come before profit-making interests. There is still a long way to go for a corporate sector that is becoming increasingly concerned. Private companies are not used to answering questions or disclosing their processes. The corporate culture *prioritises* product development and product launches without sufficient attention to ethical implications. In addition, the technology sector is becoming increasingly competitive and keeping product and service information protected by trade secrecy ensures that companies and their competitive advantages are kept away from direct imitation.

However, when these ethical recommendations become mandatory rules, a different picture emerges. Transparency now takes the form of concrete requirements, such as the labelling of artificially generated content and the publication of more information about the data on which a base model has been trained[33], for example. The problem remains, however, when compliance with them and their scope depends on the state of the art, and the latter is basically defined by the same small group of companies to which the rules apply.

Let us look at some concrete examples. The EU AIA sets out a number of requirements for all foundational, base or general models. However, it adds some additional requirements for more powerful AI systems based

[30] Kang Cecilia, "OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing" *The New York Times*, (2023), available at: https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html (accessed 1 March 2024).

[31] "Why Google CEO calls for better regulation of Artificial Intelligence" *Mix* (2023), available at: https://gestion.pe/mix/gente/inteligencia-artificial-google-ceo-pide-que-se-regule-la-inteligencia-artificial-openai-chatgpt-bard-noticia/#google_vignette (accessed 4 March 2024).

[32] Richards Blake, Aguera and Arcas Blaise, Lajoie Guillaume and Sridhar Dhanya, "The Illusion Of AI's Existential Risk" *Noema*, (2023), available at: https://www.noemamag.com/the-illusion-of-ais-existential-risk/ (accessed 1 March 2024).

[33] Heikkilä Melissa "The Five Keys to the EU's Artificial Intelligence Law" *MIT Technology Review*, (2023), available at: https://www.technologyreview.es/s/15997/las-cinco-claves-sobre-la-ley-de-la-inteligencia-artificial-de-la-ue (accessed 1 March 2024).

on the computational power needed to train them. While it is not known whether the limit would include models such as GPT-4 or Gemini, only the developers themselves know how much computing power they used to train their models. As a European Commission official acknowledged, while the technology develops, the way its power is measured and recognised should be changed[34], also to make it more transparent.

Something similar may be the case for human oversight. The European regulation dedicates Article 14 to effective human oversight, which shall aim to prevent or reduce the risks of AI systems categorised as high risk. While it is clarified in paragraph 3 that this shall be ensured by measures that are technically feasible, there is a risk that in some cases this possibility may become obsolete, especially in dynamic and highly complex environments.

To exemplify, even in the natural language domain, tasks can already be identified that are significantly difficult for a human to supervise. Imagine the ability to summarise text, where the evaluator must have a deep knowledge of both the text being summarised and the summarised text. This requires the human to spend a great deal of time paying close attention, making it easy for errors to be made. As if that were not enough, the example must be repeated a considerable number of times, as one piece of writing is not enough to assess a skill.

In response to this impractical approach, technology companies and researchers have developed automated methods of evaluation. This involves one AI system scrutinizing another. But if the intention is still to make human monitoring possible —to achieve the right approach to alignment and other valuable purposes that justify it— approaches continue to be devised to make this effective, convenient, and possible, at least to some extent.

For example, OpenAI has created a debate test, in which two artificial agents discuss a topic with each other and the human judges the exchange. Even if these systems have a more advanced understanding of the problem than the judge, the human may be able to judge which agent has the better argument (similar to expert witnesses arguing to convince a jury)[35]. Another possibility that has been postulated is to decompose highly intricate tasks that humans could neither judge nor perform, such as designing a complicated

[34] Heikkilä Melissa "The Five Keys to the EU's Artificial Intelligence Law" *MIT Technology Review*, (2023), available at: https://www.technologyreview.es/s/15997/las-cinco-claves-sobre-la-ley-de-la-inteligencia-artificial-de-la-ue (accessed 1 March 2024).

[35] Amodei Dario, Irving Geoffrey, "AI safety via debate", *OpenAI Blog*, (2018), available at: https://openai.com/research/debate (accessed 2 March 2024).

transit system or managing every detail of the security of a large computer network, into smaller subtasks or components that can be assessed[36].

3. Third. A risk-based approach

If there is one aspect of the European regulation that is already causing the famous "Brussels effect", it is the approach based on risk identification. Countries such as Australia and Canada have also drawn the line between high-risk systems in order to impose stricter requirements and obligations. Broadly speaking, this involves creating categories of systems according to their risk, and assigning compliance obligations to each of them. Its benefit is that it allows for early regulatory intervention focused on preventing harm while imposing commensurate costs with potential negative impacts.

This trend is even promoted by international cooperation groups. The G7 member countries (Canada, France, Germany, Italy, Japan, the United Kingdom, the United States and the EU) expressed a unified vision on AI and called for AI policies to be risk-based[37]. They also reached agreement on International Guiding Principles on AI and on a Code of Conduct for developers.

First, it should be clarified that the proposed EU AIA contains *ex-ante* obligations to ensure security, cybersecurity and the protection of fundamental rights, as well as *ex-post* liability rules to compensate for damages when an AI risk materialises[38]. The risk-based approach is placed in the first group of preventive measures.

In particular, the AIA classifies AI tools into different groups: unacceptable risk systems such as certain types of biometrics, which it prohibits; high risk systems that could have an adverse impact on security or fundamental rights such as recruitment systems, which will have to comply with a number of specific requirements; limited risk systems such as chatbots that will be subject to transparency rules; and finally, minimal risk systems for which voluntary measures will be put in place.

On the other hand, the US has also adhered to a risk-based regulatory

[36] Christiano Paul, Amodei Dario, "Learning complex goals with iterated amplification" *OpenAI*, (2018), available at: https://openai.com/research/learning-complex-goals-with-iterated-amplification (accessed 2 March 2024).

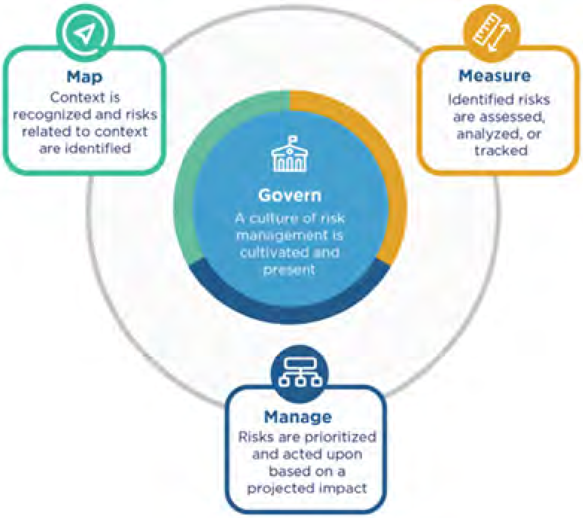
[37] Morini Bianzino et. al, "The Artificial Intelligence (AI) global regulatory landscape. Policy trends and considerations to build confidence in AI" *EY,* (2024).

[38] Kretschmer Martin, Kretschmer Tobias, Peukert Alexander, Peukert Christian, "The risks of risk-based AI regulation: taking liability seriously", *ArXiv:2311.14684v1* (2023).

approach. President Biden's Executive Order[39] mentions addressing AI risks in several opportunities. However, it does not define categories or criteria for classification.

In this regard, the National Institute of Standards and Technology's (NIST) Artificial Intelligence Risk Management Framework[40], developed in collaboration of the public and private sectors and published in January 2023, can be taken as a reference. This framework is applicable as voluntary guidance for both policy makers developing risk-based AI regulations and companies considering how to organise their internal AI governance.

This Framework foresees a set of core functions for risk management: governing, mapping, measuring and managing. Each of these high-level functions is divided into categories and sub-categories, which in turn contain specific actions and outcomes. The process is expected to be carried out by a diverse team in an interactive and non-linear way. The aim is to create opportunities to surface issues and identify existing and emerging risks.


Map
Context is recognized and risks related to context are identified
Measure
Identified risks are assessed, analyzed, or tracked
Govern
A culture of risk management is cultivated and present
Manage
Risks are prioritized and acted upon based on a projected impact

Source: AI Risk Management Framework (AI RMF 1.0), NIST, 2023

[39] The Executive Order is available at: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/ (accessed 3 March 2024).

[40] The Framework is available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (accessed 3 March 2024).

III. Regulatory impulses from other parts of the world

So far, we have highlighted the main regulatory trends by providing concrete examples. We will now focus on a few states in particular, highlighting relevant issues in their regulatory framework. To complement these explanations, we have developed a comparative table summarising how different states have resolved the dichotomies presented in the previous section.

1. Australia

In recent months, the Australian government has been criticised for its lack of action around seizing opportunities and responding to the risks posed by AI[41]. While it has been slower to react than other countries, it has already taken its first steps towards regulation. Following a public consultation on safe and responsible AI that has demonstrated Australian society's desire for strong protections through over 500 responses, the government has issued an interim response in early 2024[42].

The text reports that mandatory requirements are being developed in high-risk environments only, because there the damage caused will be impossible to reverse. For these, pre-testing to ensure safety, transparency and accountability will be central. Work is also underway with industry to develop a safety standard and labelling options for AI-generated content, both of which are voluntary.

In this sense, the approach is to combine specific obligations on high-risk AI with a voluntary "soft law" of lighter touch for less risky uses[43]. The aim is to achieve a balance that demonstrates to citizens an active response to their concerns and the intention to protect consumers, in line with international developments, but which also manages to encourage AI adoption and innovation through close collaboration with industry.

[41] Taylor Josh, "Australia "at the back of the pack" in regulating AI, experts warn" *The Guardian*, (2023), available at: https://www.theguardian.com/australia-news/2023/nov/07/australia-ai-artificial-intelligence-regulations-back-of-pack (accessed 3 March 2024).

[42] The document is available on the official website of the Australian Government Department of Industry, Science and Resources: https://consult.industry.gov.au/supporting-responsible-ai (accessed 3 March 2024). Also for a summary see: "*Action to help ensure AI is safe and responsible*" *Minister of Industry and Science*, (2024), available at: https://www.minister.industry.gov.au/ministers/husic/media-releases/action-help-ensure-ai-safe-and-responsible (accessed 3 March 2024).

[43] Lincoln Julian, Wilkinson Susannah, Lundie Alex, "Australian Government announces mandatory regulation for high-risk AI" *Herbert Smith Freehills*, (2024), available at: https://www.herbertsmithfreehills.com/insights/2024-01/australian-government-announces-mandatory-regulation-for-high-risk-AI (accessed 3 March 2024).

On the other hand, the practical impact of the proposal remains unclear given the uncertainty of the scope of the definition of high-risk AI uses. Two examples were given in the consultation document: robots used in surgery and autonomous vehicles. This definition will be considered during further consultation along with the obligations that may be imposed, so businesses that consider themselves affected should consider active participation.

Moving away from future plans and focusing on existing legislation, Australia has shown a strong commitment to some sectoral regulations. For example, on Generative AI, it has issued a Technology Trends Position Statement on Generative Artificial Intelligence, a document that examines the landscape of this technology, identifying examples of misuse and potential risks. It also reviews regulatory challenges and approaches, establishing that the eSafety commissioner -the office who issues the document- uses a multi-faceted approach involving prevention, protection and proactive, systemic change[44].

Also, in relation to the latter type of AI, at the federal level there is an Australian Framework for Generative Artificial Intelligence in Schools[45] and an Interim Guidance for Government Use of Public Generative Artificial Intelligence Tools[46]. On the other hand, at state level there is the New South Wales Government's Basic Guide to Generative Artificial Intelligence[47] and the Queensland Government's Guide to the Use of Generative Artificial Intelligence[48].

Another sectoral example of AI-related regulation that is being worked on intensively is the case of autonomous vehicles and the various documents and analyses of the Australian National Transport Commission[49]. One of the

[44] The document is available at: https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf (accessed 3 March 2024).

[45] The Australian Framework for Generative AI in Schools is available at: https://www.education.gov.au/schooling/resources/australian-framework-generative-artificial-intelligence-ai-schools (accessed 3 March 2024).

[46] The Interim Guidance for Government Use of Generative AI Public Tools is available at: https://architecture.digital.gov.au/guidance-generative-ai (accessed 3 March 2024).

[47] The New South Wales Government's Basic Guide to Generative Artificial Intelligence is available at: https://www.digital.nsw.gov.au/policy/artificial-intelligence/generative-ai-basic-guidance (accessed 3 March 2024).

[48] The Queensland Government's Guidance on the Use of Generative Artificial Intelligence is available at: https://www.forgov.qld.gov.au/information-and-communication-technology/qgea-policies-standards-and-guidelines/use-of-generative-ai-in-queensland-government (accessed 3 March 2024).

[49] Other documents that can be mentioned are: the 2017 "*National Law Enforcement Guidelines for Automated Vehicles*", available at: https://www.ntc.gov.au/sites/default/files/assets/

most recent is the Autonomous Vehicles Regulatory Framework issued in 2022[50] which presents final proposals for amendments to current legislation to accommodate the commercial use and deployment of driverless vehicles.

2. Canada

Part Three of the Digital Charter Bill C-27 would implement the Artificial Intelligence and Data Act (AIDA) to regulate the responsible development of AI in the Canadian marketplace[51]. This regulation follows the global trend by adopting a risk-based approach. It actually regulates "high-impact" systems whose specific accuracy and requirements will be developed after consultation with stakeholders, a process similar to that proposed by Australia.

In reality, regulation is planned to define the criteria for identifying high-impact AI systems, so that upgrades can occur more nimbly as the technology advances. That's because the benefits and risks of AI are still emerging, and even technology experts cannot predict where the AI market will go next.

The government currently considers that these are key factors: 1. Risks of harm to health, safety or human rights, based on both the intended purpose and the potential unintended consequences; 2. The severity of the potential harms; 3. The scale of use; 4. The nature of the harms or adverse impacts that have already occurred; 5. The extent to which, for practical or legal reasons, it is not reasonably possible to opt out of that system; 6. Imbalances in economic or social circumstances, or the age of those affected; and; 7. The level in which the risks are adequately regulated by other law.

It is also advanced that the obligations of high-impact AI systems would be guided by the principles of: human oversight and monitoring, transparency, fairness, security, accountability, validity and robustness.

In addition, the AIDA provides for several types of sanctions for

files/AV_enforcement_guidelines.pdf (accessed 3 March 2024) and the 2018 "*Discussion Paper on Motor Vehicle and Automated Vehicle Injury Insurance*", available at: https://www.ntc.gov.au/sites/default/files/assets/files/NTC%20Discussion%20Paper%20-%20Motor%20Accident%20Injury%20Insurance%20and%20Automated%20Vehicles.pdf (accessed 3 March 2024).

[50] The *Autonomous Vehicle Regulatory Framework* issued in 2022 is available at: https://www.ntc.gov.au/sites/default/files/assets/files/NTC%20Policy%20Paper%20-%20regulatory%20framework%20for%20automated%20vehicles%20in%20Australia.pdf (accessed 3 March 2024).

[51] Details of the standard are available at: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document#s10 (accessed 3 March 2024).

non-compliance. First, administrative monetary penalties that could be imposed by the regulator for any violation in order to encourage compliance. Second, prosecution of regulatory violations, which are foreseen for more serious cases where guilt must be proven beyond reasonable doubt. Finally, a separate mechanism for criminal offences, where the prohibition is violated by conscious or intentional behaviour and serious harm is caused.

While work on this project continues, the Minister of Innovation, Science and Industry announced the "Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative Artificial Intelligence Systems"[52], which temporarily provides Canadian companies with common standards and allows them to demonstrate, on a voluntary basis, that they are developing and using these systems responsibly until formal regulation is in place. The aim is to strengthen public confidence in these systems.

IV. A Latin American perspective

Regarding Latin American countries, as we have already mentioned, in recent years, AI regulation efforts have focused on ethical recommendations, strategic plans and updates to personal data protection ecosystems. However, there were draft laws that consisted of comprehensive regulatory attempts, in the countries of Mexico, Chile and Brazil. These are described below.

1. Mexico

Deputy Ignacio Loyola presented a first bill to establish a legal framework around the use and development of AI in the country. It would be called, "Law for the Ethical Regulation of Artificial Intelligence and Robotics" and would aim to regulate the use of this technology for governmental and economic purposes so that it is always based on ethics and law. The proposal is to create a bureaucratic body called the "Mexican Council of Ethics for Artificial Intelligence and Robotics (CMETIAR)" that will serve as a platform where professionals from different sectors will come together to develop and propose new rules[53].

[52] The Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative Artificial Intelligence Systems is available at: https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems (accessed 3 March 2024).

[53] González Fernanda, "Presentan propuesta de ley para regular la AI en México",

This new entity would also be made up of a representative of the executive branch appointed by the President of Mexico, as well as members of the National Council of Humanities, Science and Technology, the National Human Rights Commission (NHRC) and the Congress of the Republic. It is also made up of civilians and some players from the private sector[54]. Its function will be to review ethical protocols, oversee compliance with the norms and deliver reports on this monitoring.

Finally, the use of AI and Robotics for purposes of social manipulation, discrimination or violation of the rule of law are prohibited practices.

However, in the last days of February 2024, Senator Ricardo Monreal presented in the Senate Gazette the Initiative with a draft decree that would enact the Federal Law Regulating Artificial Intelligence[55]. It has 25 articles and adopts a risk approach similar to that of the European Union. The main aspects of its approach have been analysed in the comparative table accompanying this article.

In addition to these projects, Mexico has Recommendations for the treatment of personal data derived from the use of Artificial Intelligence (2022)[56] and a National AI Agenda (2020)[57]. The latter has a chapter on ethics that addresses the following issues: freedom of expression, privacy, equality and non-discrimination, human rights and democracy. It is worth noting that a specific analysis is dedicated to minority groups such as indigenous peoples and women on issues such as the digital divide. Finally, recommendations are made to the different actors, such as maintaining channels of dialogue with citizens, the creation of regulatory sandboxes and the measurement of risks.

The Wired, (2023), available at: https://es.wired.com/articulos/diputado-presenta-propuesta-de-ley-para-regula-la-AI-en-mexico (accessed 2 March 2024).

[54] González Fernanda, "Presentan propuesta de ley para regular la AI en México", *The Wired*, (2023), available at: https://es.wired.com/articulos/diputado-presenta-propuesta-de-ley-para-regula-la-AI-en-mexico (accessed 2 March 2024).

[55] Riquelme Rodrigo, "Ricardo Monreal presenta iniciativa para regular la inteligencia artificial" *El Economista*, (2024), available at: https://www.eleconomista.com.mx/tecnologia/Ricardo-Monreal-presenta-iniciativa-para-regular-la-inteligencia-artificial-20240227-0055.html (accessed 18 March 2024).

[56] The document Recommendations for the processing of personal data arising from the use of Artificial Intelligence is available at: https://home.inai.org.mx/wp-content/documentos/DocumentosSectorPublico/RecomendacionesPDP-AI.pdf (accessed 9 March 2024).

[57] Mexico's National AI Agenda document is available at: https://36dc704c-0d61-4da0-87fa-917581cbce16.filesusr.com/ugd/7be025_6f45f669e2fa4910b32671a001074987.pdf (accessed 9 March 2024).

2. Chile

In April 2023, a bill on robotics, Artificial Intelligence and related technologies was introduced in the Chilean Congress that seeks to regulate AI systems. This legislative initiative makes a risk-based classification of AI systems, dividing them into unacceptable risk systems and high-risk systems, similar to the European approach.

On the other hand, Chile has a National Artificial Intelligence Policy[58], launched by the Ministry of Science, Technology, Knowledge and Innovation, which contains a chapter on ethics, legal and regulatory aspects and socio-economic impacts. It sets out the following objectives:

1. Promote the construction of regulatory certainties on AI systems that allow for their development, respecting fundamental rights in accordance with the Constitution and the law.

2. Promoting algorithmic transparency.

3. Conduct forward-looking analysis to actively identify the most vulnerable occupations, anticipate the creation of new AI jobs and support workers in the transition to new occupations, minimising their personal and family costs.

4. Provide support to workers in the face of automation.

5. Promote the use of AI in e-commerce that is transparent, non-discriminatory and respectful of personal data protection rules.

6. Promote an up-to-date Intellectual Property system capable of fostering and strengthening creativity and AI-based innovation, rewarding creators and innovators in a way that encourages them to make their creation and innovation public so that society as a whole can benefit from it.

7. Position AI as a relevant component in the field of cyber security and cyber defence, promoting secure technological systems.

8. Encourage the participation of women in AI-related research and development to a level equal to or higher than the OECD.

9. Promote the participation of women in AI areas in industry to at least at or above the OECD average and ensure that the impact of automation is gender neutral and that job creation is equitable.

10. Promote gender equity in the implementation of AI systems.

[58] The National Artificial Intelligence Policy is available at: https://www.minciencia.gob.cl/uploads/filer_public/bc/38/bc389daf-4514-4306-867c-760ae7686e2c/documento_politica_AI_digital_.pdf (accessed 9 March 2024).

3. Uruguay

In 2020 Uruguay[59] approved the Artificial Intelligence Strategy for the Digital Government[60], with the aim of promoting and strengthening the responsible use of AI in its Public Administration.

That document lists nine general principles that should guide the design, development and deployment of intelligent systems:

1. Purpose: Intelligent systems should enhance human capabilities, complement them and improve people's quality of life.

2. General interest: intelligent systems driven by the State should serve the general interest, ensure inclusiveness and equity, reduce unwanted biases in data and models, and not engage in discriminatory practices.

3. Respect for Human Rights: Intelligent systems must respect human rights, individual freedoms and diversity.

4. Transparency: Intelligent systems used in the public sector must be transparent and comply with current regulations, for which the algorithms and data used for their training and implementation must be made available, as well as the tests and validations carried out, and all processes that use AI, whether as support or for decision-making, must be made explicitly visible.

5. Accountability: Intelligent systems must have a responsible person clearly identifiable who is responsible for the consequences arising from the actions of the solutions.

6. Ethics: where intelligent systems present ethical dilemmas, these should be addressed and resolved by humans.

7. Added value: intelligent systems should only be used when they add value to a process. AI should not be an end in itself.

8. Privacy by design: Intelligent systems must consider the privacy of individuals by design.

9. Security: Intelligent systems must comply with basic information security principles from their design.

[59] The analysis of the ethical principles contained in the documents issued by the Uruguayan government is part of a research project carried out by UBA IALAB focused on the ethical principles of AI that have been developed by different national governments as well as by the main industry players and international organisations. See: Sánchez Caparrós Mariana, "Principios éticos para una inteligencia artificial antropocéntrica: consensos actuales desde una perspectiva global y regional" in Corvalán Juan G. (director) "Tratado de Inteligencia Artificial y Derecho" *Thompson Reuters La Ley*, (2023), 2nd edition.

[60] The Artificial Intelligence Strategy for Digital Government is available at: https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/comunicacion/publicaciones/estrategia-inteligencia-artificial-para-gobierno-digital/estrategia (accessed 9 March 2024).

4. Brazil

In Brazil there is both a draft law to regulate AI (Bill 2338/2023)[61] and a national strategy. Regarding the former, the main aspects of its approach have been analysed in the comparative table accompanying this article. However, the peculiarity of the regulation that escapes segmentation and therefore we add here is the provision on civil liability.

The latter establishes that the provider or operator of the Artificial Intelligence system that causes material, moral, individual or collective damage is obliged to repair it in full, regardless of the degree of autonomy of the system: a) When it is an Artificial Intelligence system of high risk or excessive risk, the provider or operator is objectively liable for the damage caused, to the extent of its participation in the damage; b) When it is not an Artificial Intelligence system of high risk or excessive risk, the fault of the tortfeasor will be presumed, applying the reversal of the burden of proof in favour of the victim.

Another interesting aspect is the provision for regulatory sandboxes that must meet the following requirements: a) innovation in the use of the technology or in the alternative use of existing technologies; b) improvements towards efficiency gains, cost reduction, increased safety, risk reduction, benefits for society and consumers, among others; c) discontinuity plan, with foreseen measures to be taken to ensure the operational viability of the project once the regulatory authorisation period of the sandbox has ended.

On the other hand, in line with OECD guidelines, Brazil's Strategy for Artificial Intelligence[62], published in 2021, is based on the five principles for responsible AI defined by the OECD, which must be followed at all stages of

[61] The main aspects of this project have been analysed in a recent paper published by UBA IALAB. See Corvalán Juan G. (direction), Sánchez Caparrós Mariana, Rabán Melisa (coordination), Stringhini Antonella, Papini Carina Mariel, Heleg Giselle, Bonato Valentín, "Propuestas de regulación y recomendaciones de inteligencia artificial en el mundo. Síntesis de principales aspectos" *IALAB UBA*, (2023), available at: https://ialab.com.ar/wp-content/uploads/2023/08/Propuestas-de-regulacion-y-recomendaciones-de-AI-en-el-mundo-1.pdf (accessed 9 March 2024).

[62] The analysis of the ethical principles contained in the documents issued by the Brazilian government is part of a research project carried out by UBA IALAB focused on the ethical principles of AI that have been developed by different national governments as well as by the main industry players and international organisations. See: Sánchez Caparrós Mariana, "Principios éticos para una inteligencia artificial antropocéntrica: consensos actuales desde una perspectiva global y regional" in Corvalán Juan G. (director) "Tratado de Inteligencia Artificial y Derecho" *Thompson Reuters La Ley*, (2023), 2nd edition.

AI development and use and may even be elevated to normative requirements for all government initiatives in the field.

The principles referred to in the document are:

1. Inclusive growth and development, sustainable development and well-being: smart systems must benefit people and the planet, drive inclusive growth and sustainable development, as well as well-being.

2. Human-centred values and equity: AI must respect the rule of law, human rights, democratic values and diversity, and include appropriate safeguards to ensure a fairer society.

3. Transparency and explainability: transparency and responsible disclosure in relation to intelligent systems must be ensured in accordance with the rules of the art, which allow for promoting a general understanding of these systems, that people are aware of when they interact with AI, that those affected can understand how the outcome has been produced and that those adversely affected can question it.

4. Robustness, safety and security: systems must be robust and secure throughout their lifecycle, and potential risks must be continuously managed.

5. Accountability and responsibility: depending on the application of AI and the associated risks, governance structures must be put in place to ensure the adoption of the principles for reliable AI and the mechanisms for their enforcement. The idea of accountability should be guided by the precautionary principle.

6. Transparency and explanation: systems must be able to provide meaningful and understandable information -without compromising the confidentiality of the model- about their design, operation and impact, both to developers and to users and individuals who may be affected by their decisions and outcomes.

7. Privacy: systems should respect the privacy of individuals, avoid using unauthorised information and profiling.

8. Human control of decisions: when dealing with systems with relative autonomy in decision making, humans should always be in control, especially in the implementation stage. Human control should be proportional to the level of risk of the systems.

9. Security: systems must not violate the integrity and physical and mental health of individuals. The security and confidentiality of personal data, especially sensitive data, is essential to avoid affecting the physical and mental security of individuals.

10. Responsibility: this must be based on the solidarity of the various actors involved in the life cycle of the systems for the damage that their use may cause to people.

11. Non-discrimination: systems cannot have outcomes or responses that affect the rights of specific groups or historically marginalised populations; they must adopt a gender-neutral approach and ensure that this parameter is not used as a discriminatory factor; they cannot be limited to a specific group on the basis of race, sex, religion, age, disability or sexual orientation.

12. Inclusion: Historically marginalised groups should be involved in the life cycle of systems, as well as in their evaluation.

13. Prevalence of children's rights: Artificial Intelligence systems must recognise, respect and prioritise the rights of children and adolescents; they must always respect their best interests; they must empower and educate them so that they can take an effective part.

14. Social benefit: systems should enable or be directly related to an activity that generates a clear and identifiable social benefit. Those that pursue other purposes should not be implemented in the public sector and their use in other sectors should be discouraged.

5. Colombia

Colombia[63] is one of the Latin American states that has issued more documents to address different aspects of AI, demonstrating its commitment to the creation of a responsible ecosystem. On the one hand, there are national strategies, agendas and plans: Strategic Plan for the Transfer of Knowledge in Artificial Intelligence[64], Plan for Monitoring the Implementation in Colombia of International Principles and Standards in Artificial Intelligence[65] and

[63] The analysis of the ethical principles contained in the documents issued by the Colombian government is part of a research project carried out by UBA IALAB focused on the ethical principles of AI that have been developed by different national governments as well as by the main industry players and international organisations. See: Sánchez Caparrós Mariana, "Principios éticos para una inteligencia artificial antropocéntrica: consensos actuales desde una perspectiva global y regional" in Corvalán Juan G. (director) "Tratado de Inteligencia Artificial y Derecho" *Thompson Reuters La Ley*, (2023), 2nd edition.

[64] The document Strategic Plan for Knowledge Transfer in Artificial Intelligence is available at: https://dapre.presidencia.gov.co/AtencionCiudadana/DocumentosConsulta/consulta-Plan-estrategico-transferencia-conocimiento-inteligencia-artificial-210708.pdf?TSPD_101_R0=08394a21d4ab2000ad47a40d2942398a3afd43b1cf6ddc68ee01a62e6b7ddba4ba90e5fef6630a4608e2a81956143000a21c50a82d22231f3752d884d7f114087af3c80c0db6ca300c0a7476cfb73e4532ed193a19e700d58d63817dba9c2eae (accessed 9 March 2024).

[65] The document Plan de Seguimiento a la Implementación en Colombia de Principios y Estándares Internacionales en Inteligencia Artificial is available at: https://dapre.presidencia.gov.co/TD/plan-seguimiento-implementacion-colombia-estandares-internacionales-inteligencia-artificial-ocde.pdf?TSPD_101_R0=08394a21d4ab20003ce781987b45f801b436fefee21570395b2f0af80498840c752d7f9356e396f508f3d002e214500049b04c4c1af8bc686cdc-

the National Policy for Digital Transformation and Artificial Intelligence[66] , among others. On the other hand, criteria for regulatory sandboxes have also been issued, such as the Conceptual Model for the Design of *Regulatory Sandboxes & Beaches* in Artificial Intelligence (draft document for discussion)[67] and the Sandbox on privacy by design and by default in Artificial Intelligence projects[68].

In October 2021 Colombia approved its Ethical Framework for Artificial Intelligence, with the aim of protecting, strengthening and guaranteeing human rights in the development, use and governance of AI. This document recognises the following ethical principles.

6. Argentina

The Undersecretary for Information Technologies approved in June 2023 the "Recommendations for Reliable Artificial Intelligence"[69]. With this initiative, the country seeks to ensure the responsible and beneficial development of AI by strengthening the scientific and technological ecosystem.

The document, which is about 30 pages long, is aimed primarily at those in the public sector and elaborates specific ethical recommendations to consider at each stage of the life cycle of these technologies. For example, before starting, it is appropriate to assemble a diverse, multidisciplinary team and explore other types of less costly technologies that can solve the problem. Then, data issues such as privacy, quality and validation, among others, are addressed.

6b0aedc6392a3f57fcc1b8445a48cb55659b6841af5a10357db7c1294aa242aefd7aa3202b95e-19da334e85bbf489163be0308e025c7655769e9ae6b38f2593551645e60ed63 (accessed 9 March 2024).

[66] The National Policy for Digital Transformation and Artificial Intelligence is available at: https://colaboracion.dnp.gov.co/CDT/Conpes/Económicos/3975.pdf (accessed 9 March 2024).

[67] El documento se encuentra disponible en: https://dapre.presidencia.gov.co/AtencionCiudadana/DocumentosConsulta/consulta-200820-MODELO-CONCEPTUAL-DISENO-REGULATORY-SANDBOXES-BEACHES-AI.pdf?TSPD_101_R0=08394a21d4ab20003a42e18525bea92cf2a46d01179eb2f6f8cce49d0b07777314d10d-54b571ead00826b165171430002696a9a61e05251cfe9fbf7caef2d41ace97aad23d8712bd8f-78dabd992658305ff1241c103fa8683ef189469120601c (consultado el 9 de marzo de 2024).

[68] The document is available at: https://www.sic.gov.co/sites/default/files/normatividad/112020/031120_Sandbox-sobre-privacidad-desde-el-diseno-y-por-defecto.pdf (accessed 9 March 2024).

[69] The Recommendations for Trusted Artificial Intelligence are available at: https://www.boletinoficial.gob.ar/detalleAviso/primera/287679/20230602 (accessed 9 March 2024).

Argentina thus joins international efforts on AI ethics, taking into account precedents such as the UNESCO Recommendation on the Ethics of Artificial Intelligence, the Asilomar Conference, the OECD Council of Ministers meetings and the G20 Ministerial Meeting on Trade and the Digital Economy. In doing so, it collects and analyses all the ethical principles contained in these documents or elaborated by these groups.

The document identifies the human-centred AI approach with the requirement that the respective actors observe the rule of law, human rights and democratic values throughout the life cycle of the AI system. These values include freedom, dignity and autonomy, privacy and data protection, non-discrimination and equality, diversity, equity, social justice and internationally recognised labour rights.

It also mentions two types of models that can be chosen to adopt AI. One is automation, in which human intervention is limited to the control of the system. This paradigm is the one most associated with the idea of replacing human capabilities with the new functions of intelligent systems. The second is the *human-in-the-loop* paradigm, which involves human-machine collaboration to solve problems. This selectively includes the participation of people, in order to take advantage of the benefits or the most efficient aspects of both components leading to an augmented intelligence solution.

Finally, it is worth mentioning some aspects of the recommendations made at the "Conference on regulation and legislation of Artificial Intelligence: Generative AI and international trends" held on 5 June 2023 in the Chamber of Deputies of Argentina by the UBA IALAB team[70]. There, it was considered advisable to deal with a general and broad regulatory framework as a law of minimum standards to avoid anachronisms, that include ethical principles and provide for a measurement of the possible impact of AI and, due to its particularity, the approach to the risks that it entails.

The key was found in designing a law (or body of laws) that allows all actors involved in the AI life cycle to make autonomous, conscious and informed decisions. In this sense, it was considered that dispersed and implicit regulations are detrimental to visibility, so it is advisable to enact rules on the basis of minimum standards that explicitly uphold principles, rights and obligations[71].

[70] On the consensus and recommendations resulting from the conference, see: "Puntos de partida para la regulación de la inteligencia artificial en Argentina" in Corvalán Juan G. (director), "Tratado de Inteligencia Artificial y Derecho" *Thompson Reuters La Ley*, (2023), 2nd edition.

[71] On the consensus and recommendations resulting from the conference, see: "Puntos

V. Conclusion

On the one hand, in terms of the regional situation in Latin American countries, there has been a progressive advancement in the creation of regulatory frameworks. Strategic plans for AI coexist with ethical guidelines that seek to guide its responsible development. For these documents, the recommendations drawn up by international organisations such as the OECD serve as points of reference, consistent with the global trend.

Globally, despite being at different points in their development and with different economic and social circumstances, countries follow very similar trends in AI regulation. The greatest variability can be observed in the mandatory or voluntary nature of the different frameworks and the extent to which the promotion of innovation becomes a priority to be taken into account when making this decision.

This intuition is consistent with a study conducted by the consultancy Deloitte, which analysed a database of more than 1,600 AI policies ranging from regulations to research grants and national strategies[72]. Rather than finding clear sets of related policies, it was found that most policies were lumped together. It was also revealed that there is not only overlap in the basic policies, but also in the path countries follow towards regulation. Almost all countries follow the path of understanding the technology, growing and stimulating the industry and then shaping it through regulatory instruments and standards.

This common approach suggests a convergence in the global understanding of the challenges and opportunities presented by AI, as well as the need for a regulatory framework that encourages innovation while ensuring safety and ethics in its application.

de partida para la regulación de la inteligencia artificial en Argentina" in Corvalán Juan G. (director), "Tratado de Inteligencia Artificial y Derecho" *Thompson Reuters La Ley*, (2023), 2nd edition.

[72] The study is available at: Mariani Joe, Eggers William D., Kamleshkumar Kishnani Pankaj, "The AI regulations that aren't being talked about*" Deloitte,* available at: https://www2.deloitte.com/xe/en/insights/industry/public-sector/ai-regulations-around-the-world.html (accessed 4 March 2024).

“Artificial Intelligence”, territorial scope and scope of the Regulation and its relationship with data protection

WHAT IS "ARTIFICIAL INTELLIGENCE" FOR THE REGULATION? ANALYSIS, DELIMITATION AND PRACTICAL APPLICATIONS

Lorenzo Cotino Hueso

Professor of Constitutional Law at the University of Valencia. Valgrai

I. The importance of the concept of Artificial Intelligence in the Regulation

The AIA concept of AI is key,[1] inter alia, because it essentially determines the application of the AIA, which revolves around AI, high-risk AI systems, certain AI systems, general-purpose AI models or AI systems already introduced in the market. Hence, legally delimiting the notion of AI is the essential premise.

Thus, it should be recalled that Article 1 regulates the "purpose" of the regulation ("to promote the adoption of human-centred and trustworthy Artificial Intelligence" and revolves around "AI systems"). Thus, the AIA establishes harmonised rules, prohibitions, specific requirements for high-risk AI systems, market monitoring and surveillance mechanisms, as well as innovation tools (Art. 1.2). Article 2 on the "scope of application" focuses on the various subjects in the value chain of "AI systems" and high-risk systems. In addition, reference is made to "AI systems or models" in research (Art. 2. 6th and 8th). Thus, the concepts of AI system and "high risk" determine the application of the standard. This is, of course, without prejudice to the importance of "AI models for general use" (Art. 1. 2º e) and 2. 1º a), Chapter V) or the transparency obligations of "certain AI systems" (Art. 50, Chapter IV).

Where appropriate, reference should be made to the Final Provisions (Ch. XIII) in relation to "AI systems already placed on the market or put into service" (Art. 111). Here reference is made to "AI systems which are compo-

[1] cotino@uv.es. OdiseIA. This study is the result of research from the following projects: MICINN Project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by MCIN/AEI/10.13039/501100011033/; "The regulation of digital transformation ..." Generalitat Valenciana "Algorithmic law" (Prometeo/2021/009, 2021-24); "Algorithmic Decisions and the Law: Opening the Black Box" (TED2021-131472A-I00) and "Digital transition of public administrations and Artificial Intelligence" (TED2021-132191B-I00) of the Recovery, Transformation and Resilience Plan. CIAEST/2022/1, Research Group in Public Law and ICT, Catholic University of Colombia; MICINN; CIAEST/2022/1, Digital Rights Agreement-SEDIA Scope 5 (2023/C046/00228673) and Scope 6 (2023/C046/00229475).

nents of large-scale IT systems" as regards the application of the AIA. These systems are identified in Annex X and relate to the area of freedom, security and justice (Schengen, Visas, Eurodac, Criminal Records, etc.).

II. The various definitions of Artificial Intelligence have been shuffled

It is not a simple task to define what Artificial Intelligence is. Since the 20th century, more than 55 definitions have been identified,[2] this can have very different perspectives such as research, policy and institutional, economic and market. It should also be noted that "Artificial Intelligence" attracts a lot of investment, so that many so-called AI systems are nothing more than the name and have little or nothing to do with the concept of AI from the AIA, which is the one to follow.

The difficulties for a definition of AI are greatest when it comes to projecting a legal regime into an AI system. A definition for legal purposes has clear political and institutional objectives and, at the same time, as much legal certainty as possible is required. In all cases, seeking a definition has the difficulty that it must be adaptive to the necessary changes that technology will bring in the future.

Thus, the European Union has followed an evolution of concepts by the European Commission in 2018[3], by the Commission's High Level Expert Group in 2019[4], by the European Parliament in

[2] *Thus,* Samoili, S., et al., AI *WATCH. Defining Artificial Intelligence*, Publications Office of the European Union, Luxembourg, 2020, doi:10.2760/382730, JRC118163. https://publications.jrc.ec.europa.eu/repository/handle/JRC118163

[3] Thus, in the Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions *on Artificial Intelligence for Europe*, Brussels, 25.4.2018 COM(2018) 237 final, p. 1, it was stated that "Artificial Intelligence (AI) refers to systems that display intelligent behaviour by analysing their environment and taking actions, with a certain degree of autonomy, to achieve specific goals. AI-based systems can be purely software-based, acting in the virtual world (e.g. voice assistants, image analysis software, search engines, voice and face recognition systems) or AI can be embedded in hardware devices (e.g. advanced robots, autonomous cars, drones or Internet of Things applications)".

[4] European Commission Independent High Level Expert Group on Artificial Intelligence (HLEG), *A definition of AI: main capabilities and disciplines*, Definition developed for the purpose of the AI HLEG's deliverables, European Commission. 8 April 2019, p. 6. We propose to use the following updated definition of AI: "Artificial Intelligence (AI) systems are human-designed software (and possibly also hardware) systems (*) that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning about the knowledge, or

2020[5]. Finally, seeking a broader international consensus the EU for AIA opted for the OECD concept in its Recommendation of the "AI Principles" in 2019[6] . The OECD definition in 2019 was based on Russell and Norvig's 2009 concept[7] . At this point it should be noted that on 8 November 2023 the OECD Council has modified its concept of AI to align also with the final versions of the AIA, as well as the concepts of Japan and other countries. The concept is as follows: "An AI system is a machine-based system that, by

processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can use symbolic rules or learn a numerical model, and can also adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), automatic reasoning (which includes planning, scheduling, knowledge representation and reasoning, search and optimisation), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques in cyber-physical systems)".

[5] European Parliament resolution of 20 October 2020 with *recommendations to the Commission on a framework for the ethical aspects of Artificial Intelligence, robotics and related technologies* (2020/2012(INL)). ANNEX B. setting out the *Proposal for a Regulation of the European Parliament and of the Council on ethical principles for the development, deployment and use of Artificial Intelligence, robotics and related technologies*. (a) 'Artificial Intelligence' means a system, whether software-based or embedded in physical devices, that exhibits intelligent behaviour by being able, inter alia, to collect and process data, analyse and interpret its environment and take action, with a certain degree of autonomy, in order to achieve specific objectives.

[6] OECD, *Recommendation of the Council on Artificial Intelligence*, 22 May 2019, OECD/LEGAL/0449, adopted by the OECD Council at Ministerial level on 22 May 2019, https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. This document "AGREES that, for the purposes of this Recommendation, the following terms should be understood as follows:

- AI system: An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions that influence real or virtual environments. AI systems are designed to operate with different levels of autonomy.

- AI system lifecycle: The phases of the AI system lifecycle include: (i) "design, data and modelling", which is a context-dependent sequence encompassing planning and design, data collection and processing, as well as model building; (ii) "verification and validation"; (iii) "deployment"; and (iv) "operation and monitoring". These phases often take place iteratively and are not necessarily sequential. The decision to remove an AI system from operation can occur at any time during the operation and monitoring phase."

[7] It stated that "An AI system is a machine-based system that can, for a given set of human-defined goals, make predictions, recommendations or decisions that influence real or virtual environments. AI systems are designed to operate with varying levels of autonomy". OECD, *Explanatory memorandum on the updated OECD definition of an AI system*, OECD Artificial Intelligence Papers, March 2024 No. 8, Paris, https://doi.org/10.1787/623da898-en, p. 4. The reference is Russell, S. and Norvig P., *Artificial Intelligence: A Modern Approach*, 3rd edition, Pearson, London, 2009 http://aima.cs.berkeley.edu/

explicit or implicit goals, infers from the input it receives how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptability after deployment."[8] It is also worth noting that the OECD has published an interesting Explanatory Memorandum on this concept, although it is not part of the adopted text.

In the United States, a regulatory concept is dealt with in the Code, 15 U.S.C. 9401(3), reiterated in Section Two. 3. b, Executive Order on the Development and Safe and Reliable Use of Artificial Intelligence, 30 October 2023: "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions that influence real or virtual environments. Artificial Intelligence systems use machine and human-based inputs to perceive real and virtual environments; abstract those perceptions into models through analysis in an automated fashion; and use model inference to formulate choices for information or action".[9] It should be noted that there are efforts between the US and the EU to develop a common concept map and taxonomy, although the concept of AI is still pending.[10]

[8] An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment". OECD, *Explanatory memorandum... cit.* p. 4.

[9] (b) The term "Artificial Intelligence" or "AI" has the meaning set forth in 15 U.S.C. 9401(3): "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. Artificial Intelligence systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action".

https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

https://uscode.house.gov/view.xhtml?req=(title:15%20section:9401%20edition:prelim)#:~:text=(b)%20The%20term%20%22artificial,influencing%20actual%20or%20virtual%20environments

[10] Thus, following the *AI Roadmap*, Third EU-US Ministerial Declaration, the first joint roadmap on assessment and measurement tools for trustworthy AI and risk management (AI Roadmap). (AI Roadmap) https://digital-strategy.ec.europa.eu/en/library/ttc-joint-roadmap-trustworthy-ai-and-risk-management

There, a group of experts undertook to prepare an initial draft of terminologies and taxonomies of AI. Sixty-five terms have been identified with reference to key EU and US documents. However, the concept of Artificial Intelligence is "pending". See https://digital-strategy.ec.europa.eu/en/library/eu-us-terminology-and-taxonomy-artificial-intelligence

The AIA aims to provide a single definition of AI that is sufficiently clear and precise, that provides legal certainty, that is functional and as technologically neutral as possible, i.e., that does not condition or favour some technologies over others. Similarly, the aim is for a definition that will stand the test of time as well as possible given the technological and market dynamism.

III. The definition of Article 3 Regulation and its components: techniques, autonomy, adaptation, inputs and outputs, and context.

The list of 68 definitions in Article 3.1 AIA starts with "AI system": "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments". The legislative technique used for the definition of an AI system has varied throughout the legislative process. The Commission proposal referred to an Annex I. However, since the EU Council's version in December 2022, the definition is contained in full in the list of definitions, without reference to an annex. It should be noted that only in the latest version the reference to "software" is omitted.

Central to the concept of AI are the techniques used, which as the OECD recalls are varied and rapidly developing[11]. Under "AI" are included categories of techniques "such as machine learning and knowledge-based approaches, and application areas such as computer vision, natural language processing, speech recognition and intelligent decision support systems, intelligent robotic systems, as well as novel application of these tools to various domains. AI technologies are developing at a rapid pace and it is likely that additional techniques and applications will emerge in the future".[12]

Recital 12 AIA excludes from the AI concept "systems based on rules defined solely by natural persons for automatically executing operations". It is precisely in order to distinguish it from "traditional and simpler software systems or programming approaches" that the techniques are delimited.

Among the techniques that define what an AI system is is "machine learning" "inferring from coded knowledge or a symbolic representation of the task to be solved". (Cons. 12). As the OECD reminds us, "machine learning is a set

The concept paper, EU-U.S. *Terminology and Taxonomy for Artificial Intelligence. First Edition*, May 2023, p. 6 https://ec.europa.eu/newsroom/dae/redirection/document/96104

[11] OECD, *Explanatory memorandum cit.* p.1.

[12] *Ibid*, p. 6.

of techniques that allow machines to improve their performance and generally generate models in an automated way". The process of improving the performance of a system through machine learning techniques is known as "training".

Learning can be:

- supervised: based on human annotations/labelling of data.

- unsupervised: based on instances/data points that have not been tagged by a human.

- by reinforcement: based on "rewarding" the system (through trial and error), not on labelled or unlabelled datasets. Similarly, a distinction can be made between the variety of methods for machine learning and deep learning, i.e., based on neural networks (often very complex and opaque).

It is also recalled by the OECD that "machine learning can continue to adapt after the initial build phase, improving its performance by interacting directly with new inputs and data. Moreover, AI systems can be periodically upgraded/retrained, retested and redeployed as new versions".[13]

Thus, for example, deep learning can be considered to be that which uses artificial neural networks with multiple layers, such as voice recognition and image recognition systems, such as the use of convolutional neural networks to identify objects in photographs. In the case of reinforcement learning, where the system learns to make decisions through trial and error, AI systems can be placed in strategy games such as chess or *Go*, where the system improves its game through repeated play and adjusting its strategy based on the rewards obtained.

In addition to learning, reasoning or modelling techniques, which were explicitly referred to in the initial versions of AIA, should not be excluded.[14] Predictive modelling is closely related to machine learning. It includes statistical learning and inference techniques (including Bayesian estimation) and search and optimisation methods.[15] Examples include chess AI engines (search and optimisation), which generate a search tree showing some of White's possible moves.[16]

[13] *Ibid*, p. 8.

[14] Following the OECD, a model is defined as a "physical, mathematical or otherwise logical representation of a system, entity, phenomenon, process or data" in ISO/IEC 22989. It indicates that this would include, among others, statistical models and various types of input and output functions (such as decision trees and neural networks). AI models can be built manually by human programmers or automatically using, for example, unsupervised, supervised or reinforcement machine learning techniques.

[15] In previous versions it was modelling techniques, letter c of Annex 1, now there is an explicit reference in Recitals.

[16] Example of JRC modelling in Samoili, S., et al, AI *WATCH. Defining Artificial Intelligence*... cit.

An example of symbolic or knowledge-based AI systems is also given as an example of a system that reasons about manufacturing processes, which could have variables representing factories, goods, workers, vehicles, machines etc.[17]

Similarly, AI systems are reasoning-based systems that can reason (using operations such as sorting, searching, matching and chaining) on the basis of coded knowledge. These methods are more interpretable than learning systems, but can also exhibit bias, complexity, unpredictability and autonomy.[18]

Among the various techniques, we can provide examples of planning and scheduling systems that create action plans to achieve specific goals and the scheduling of tasks to be executed by a machine, such as route planning for autonomous vehicles. Regarding representation and reasoning techniques, expert systems that diagnose diseases based on the patient's symptoms and medical history can be cited.

Another substantial element of the AI concept is that the system has a minimum degree of "autonomy", in particular, "different levels of autonomy". This implies "some degree of independence" [...and] certain capabilities to function without human intervention (Cons. 12). The OECD defines autonomy as the "degree to which a system can learn or act without human involvement following human delegation of autonomy and automation of processes".[19] There are different levels of autonomy, such as the six standard levels generated in 2016 for autonomous vehicles:[20] Level 0 (no driving automation), Level 1 (driver assistance), Level 2 (partial driving automation), Level 3 (conditional driving automation), Level 4 (high driving automation) and Level 5 (full driving automation).

At the same time, "adaptive capacity" is stressed, i.e., "self-learning capabilities that allow the system to change while in use" (Cons. 12). The OECD notes that systems can continue to evolve and modify their behaviour.[21] Thus, a system can be trained and develop new forms of inference not imagined by the developers. A highly adaptive system can change its operation in response to changes in its environment, allowing it to remain effective and relevant in dynamic situations. For example, assistants such as *Siri* or *Alexa*, in their initial versions, responded to predefined commands but did not learn from

[17] *Ibid*, p. 8.

[18] Follows from *Ibid.*

[19] OECD, *Explanatory memorandum cit.* p. 6.

[20] https://www.sae.org/news/2019/01/sae-updates-j3016-automated-driving-graphic and https://www.sae.org/standards/content/j3016_202104/

[21] OECD, *Explanatory memorandum cit.* p. 6.

interactions, while advanced virtual assistants use deep learning to personalise responses and recommendations based on the user's interaction history. In the case of AI-based medical diagnostic systems, there are clinical decision support tools based on predefined rules that would not be considered AI, while AI-based diagnostic systems that use large databases and machine learning algorithms can autonomously diagnose diseases and improve their recommendations over time.

The inference capability of an AI system is also an essential defining element. On the one hand, a system generates an output from its inputs, usually after implementation. This is the "obtaining of output information, such as predictions, content, recommendations or decisions, which can influence physical and virtual environments, and the ability of AI systems to infer models or algorithms from input information or data" (Cons. 12). In terms of output types, it is noted that the categories correspond to different levels of human involvement, with 'decisions' being the most autonomous type of output (the AI system affects its environment directly or directs another entity to do so) and 'predictions' being less autonomous.[22] For example, a driver assistance system might "predict" that a region of pixels in its camera input is a pedestrian; it might "recommend" braking or it might "decide" to apply the brake. Meanwhile, generative AI systems that produce 'content' (including text, images, audio and video).

On the other hand, especially during the construction phase, the inference capability "enables learning, reasoning or modelling" (Cons. 12). Thus, the outputs of the AI system are used to evaluate a version of a model and derive a model from inputs/data.[23]

As for the input provided to the system, it may include data relevant to the task at hand or take the form of a user message or a search query. AI systems may have one or more types of goals and "may operate according to explicitly defined goals or implicit goals" (Cons. 12). Explicit goals are defined by humans. Implicit goals may be in rules (usually specified by humans) or implicit in training data. In these cases, the targets are not completely known in advance. Also, user indications may complement the targets.[24]

Finally, it should be noted that AI systems operate on machines ("machine-based") and in physical or virtual environments or contexts, and include environments that describe aspects of human activity.[25] They "can be

[22] *Ibid*, p. 8.

[23] *Idem*.

[24] *Ibid*, p. 7.

[25] *Idem*.

used independently or as components of a product, regardless of whether the system is physically part of the product (embedded) or contributes to the functionality of the product without being part of it (non-integrated)" (Cons. 12).

IV. Examples of systems that are or are not Artificial Intelligence

It must be assumed that not every computer system that enables automated processes or decisions is *automatically* AI. They will not be considered "Artificial Intelligence" even if they can reason or model mathematically. This is important; if the automated decision system is not an AI system, even if it is used for a high-risk use case and purpose (Annex III), it will also fall outside the application of the AIA.

In this direction, the JRC gives some examples.[26] A credit scoring system that aims to estimate the risk associated with granting a loan. This system uses data on borrower characteristics, financial situation (income, monthly expenses), loan amount, purpose, demographics. The result of this system is a risk category, e.g. reliable clients, clients who may have repayment problems. In this case, the requirements of deep learning architecture trained on historical information, with machine learning techniques, may be given. There may be reasoning based on a history of human decisions with logic and knowledge-based strategy techniques. There can be logistic regression on historical data with statistical strategies. However, it should not be considered as AI if the system is based on a fixed set of rules, manually defined by a human.

Another example is an algorithm for grading A Level and GCSE students in England, Wales and Northern Ireland based on historical information (past grades). [27] Thus, it will be considered to be an AI system if the relevant criteria for determining the outcome (students' grades) are chosen by humans: teachers' estimates of grades, the school's performance in previous years, the cohort's performance in previous years. In this case, in terms of the outcome approach, the statistical model combines the historical data with the teachers' estimates. The teachers' estimates are adjusted according to the statistical model to fit a distribution of past grades.

Conversely, and in the same context, an algorithm for deciding school enrolment based on student variables such as mother's level of education, eco-

[26] It follows from the JRC in Samoili, S., et al, AI *WATCH. Defining Artificial Intelligence*... cit. although such claims may be questionable.

[27] https://bera-journals.onlinelibrary.wiley.com/doi/full/10.1002/berj.3705

nomic situation, distance from home to school or student preference should not be considered as an AI system because it does not meet the requirements of Article 3. It would not be an AI system if it is humans who select the criteria that are relevant for the decision and also decide which criteria to give more importance to and what weights to assign to the categorisation criteria.

To conclude, this analysis of the concept of AI in the AIA underlines the importance of a clear and precise definition of AI to ensure legal certainty and technological functionality. The adoption of the OECD definition as the basis for AIA has been highlighted in the search for a standardised and harmonised approach at international level. The definition of AIA includes various techniques that constitute AI, such as machine learning, natural language processing and decision support systems. An attempt has been made to give a practical overview of these. In any case, the autonomy and adaptability of an AI system has been underlined. Either way, and as will be explained in the section on high-risk systems, AIA foresees evolution and adaptation to rapid technological advances.

THE TERRITORIAL SCOPE OF APPLICATION OF THE ARTIFICIAL INTELLIGENCE ACT

Alfonso Ortega Giménez

Senior Lecturer in International Private Law at the Miguel Hernández University of Elche (Alicante).[1]

I. Approach

The AIA is the first global legal regulation of Artificial Intelligence, directly applicable in all Member States of the European Union (hereinafter, EU). It is a preventive rule aimed at manufacturers/developers of AI systems so that they do not impact on the fundamental rights of individuals. At the same time, it aims to be universally effective, as has already been the case with the General Data Protection Regulation[2] , i.e., with an impact beyond the borders of the EU. It will apply to AI systems that function as components of products or are products in themselves, which are intended to be placed on the EU market or put into service (Article 2.1 of the GDPR).

This new regulation aims to develop an ecosystem of trust by establishing a legal framework to ensure that AI is trustworthy and complies with the law. It builds on the EU's fundamental values and rights, which essentially aim to inspire confidence among citizens and other users to adopt AI-based solutions, while encouraging businesses to develop and invest in such solutions. AI should be a tool for people and a positive force in society, and its ultimate goal should be to increase human well-being, while respecting people's rights.

The technique used for the regulation of this matter, which is inspired by the GDPR, is characterised by four elements[3] : a) The use of a Regulation instead of a Directive as a legal technique[4] ; b) The establishment of rigid

[1] 0000-0002-8313-2070.

[2] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of individuals with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation). OJEU *No* 119 of 4 May 2016.

[3] *See* Gascón Macén, A., "El Reglamento General de Protección de Datos como modelo de las recientes propuestas de legislación digital europea", *CDT*, Vol. 13(2) (2021), pp. 209-232, available at: https://doi.org/10.20318/cdt.2021.6256; Papakonstantinou, V. and DE HERT, P., "Post GDPR EU laws and their GDPR mimesis. DGA, DSA, DMA and the EU regulation of AI", *European Law Blog*, (2021), available at: https://europeanlawblog.eu/2021/04/01/post-gdpr-eu-laws-and-their-gdpr-mimesis-dga-dsa-dma-and-the-eu-regulation-of-ai/.

[4] Despite being referred to as "laws" in the legislative proposals. *See*, in this regard, Papakonstantinou, V. and De Hert, P., "EU lawmaking in the Artificial Intelligent Age: Actification,

requirements and obligations for different categories of positions for access to the activity and provision of any digital service; c) The appointment by the Member States of competent national authorities so that companies find a more direct way when they wish to complain about non-compliance with the Regulation; and d) The establishment of collegiate bodies at European level, although with different roles depending on each Regulation[5] .

The AIA functions as a legal tool, aiming to harmonize rules in this field and establish a robust regulatory framework that is not sector-specific. Its primary objective is to provide responses that are proportionate to the risks posed by AI. AI is designed to be used in any sector of activity, with the result that the regulatory rules of different sectors apply in relation to the design and development of AI; for example, legislation on the protection of personal data, legislation on business secrets, or legislation on consumer protection and unfair commercial practices, among others, apply[6].

The AIA is not only designed to encourage the adoption of AI systems in the internal market, but also has the ambition to position the EU as a world leader in the development of trusted and ethical AI. This legislative framework responds to the need to provide, at a global level, a high level of protection of public interests, such as health and safety, while ensuring respect for fundamental rights.

Article 2.1 of the AIA[7] becomes one of the key articles, as it delineates

GDPR mimesis, and regulatory brutality", *European Law Blog*, (2021), available at: https://europeanlawblog.eu/2021/07/08/eu-lawmaking-in-the-artificial-intelligent-age-act-ification-gdpr-mimesis-and-regulatory-brutality/.

[5] European Artificial Intelligence Committee (Article 56 AIA), although the Parliament proposes to change its name to the European Artificial Intelligence Office (AI Office) and considerably increase its functions. Other collegiate bodies foreseen in the digital laws are: European Data Innovation Board (Article 29 GDPR), European Digital Services Board (Article 61 RSD), High Level Working Party (Article 40 GDPR), which join the already existing European Data Protection Board (Article 68 GDPR).

[6] *See* Miguel Asensio, P., *Manual de Derecho de las Nuevas Tecnologías. Derecho Digital*, Aranzadi, Cizur Menor, Navarra 2023, pp. 121.

[7] Article 2.1 of the AIA states: "1. (a) providers placing on the market or putting into service AI systems or placing on the market AI models for general use in the Union, irrespective of whether those providers are established or located in the Union or in a third country; (b) deployers of AI systems that are established or located in the Union; (c) providers and deployers of AI systems that are established or located in a third country, where the output information generated by the AI system is used in the Union. (d) importers and distributors of AI systems; (e) manufacturers of products that place on the market or put into service an AI system together with their product and under their own name or trademark; (f) authorised representatives of providers that are not established in the Union; (g) affected persons that are located in the Union.

the scope of application of the law, specifying who will be subject to the new regulations; and therefore who must abide by the obligations contained in the Regulation. Providers and users of AI systems, whether within the EU or in third countries, will be affected by this Regulation when the output information from the AI system is used in the EU. This provision ensures that the regulation has a cross-border scope, covering not only actors within the EU but also those whose AI systems may affect EU citizens. The extraterritoriality of the rule needs to be carefully regulated and analysed due to the multiple implications it brings with it, being one of the major novelties of this proposal.

One of the most salient aspects of the AIA is its technology-neutral approach and its attempt to be time-resilient, taking into account the rapid evolution of AI technology and the AI market. This is essential for a durable and adaptable regulation that can keep up with technological developments without the need for frequent changes.

The AIA also provides a clear definition of the main actors in the AI value chain, such as providers[8] , deployers[9] , authorised representatives[10] , importers[11] and distributors of AI systems[12] , as well as product manufacturers who market or put into service an AI system together with their product and under their own name or trademark. This detailed approach is essential to clarify responsibilities and ensure a level playing field across the industry. On the other hand, AI systems are classified according to their ability to harm and endanger the safety and fundamental rights of individuals.

Undoubtedly, the AIA represents an ambitious effort to strike a balance between promoting technological innovation and protecting citizens and

[8] "Provider"' means a natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge.

[9] "Deployer" means a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activit.

[10] "Authorised representative" means a natural or legal person located or established in the Union who has received and accepted a written mandate from a provider of an AI system or a general-purpose AI model to, respectively, perform and carry out on its behalf the obligations and procedures established by this Regulation.

[11] "importer" means a natural or legal person located or established in the Union that places on the market an AI system that bears the name or trademark of a natural or legal person established in a third country.

[12] "Distributor" means a natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market.

their rights. What remains clear is that the EU seeks a leading position in setting global standards for AI governance, underlining its commitment to an AI system that is safe, ethical, and under human control.

Title I of the AIA defines the scope of application of the new rules covering the placing on the market by commercialisation, putting into service, and use of AI systems.

Article 2.1 of the AIA can be considered from the point of view of private international law as a rule of applicable law, and specifically, a unilateral conflict rule whose objective is to determine to which EU situations the AIA is applicable. These situations will differ depending on whether they are analysed from the position of the economic operators, the competent national authorities or the courts of law[13].

II. Application of Article 2.1 by economic operators

The first perspective of analysis is that of economic operators, i.e., "providers" placing on the market or putting into service AI systems or placing on the market general-purpose AI models in the Union, "deployers of AI systems", "authorised representatives", "importers and distributors of AI systems" and "affected persons that are located in the EU". For these operators, it is essential to know, before carrying out their economic activity, whether the Regulation will apply to them. In principle, the answer might seem simple: the future regulation will apply to AI systems that function as components of products or are themselves products that are intended to be placed on the EU market or put into service. Therefore, any AI systems developed by providers or used by users established in third States become accessible by potential customers located in Europe. Is this sufficient for the AIA to be applicable? According to Article 2.1 of the AIA, the answer must be in the negative. This provision sets out connecting criteria which, in principle, imply that it will only apply to providers and users carrying out activities which have a close link to the EU[14].

Article 2.1 of the AIA plays a crucial role in determining which economic operators will be subject to the obligations of the Regulation and how they should interpret its prospective application. The discussion on prospective application is essential for economic operators, as it provides them with the

[13] *See* López-Tarruella Martínez, A., "El reglamento de Inteligencia Artificial y las relaciones con terceros estados", *Revista Electrónica de Estudios Internacionales (REEI)*, n.º 45 (2023), pp. 5-11.

[14] *Vid.* in a broad sense, *ibid.*

necessary clarity to plan and adapt their business strategies in accordance with the new regulatory requirements.

In the context of the AIA, economic operators include providers, deployers of AI systems, product manufacturers, authorised representatives, importers and distributors of AI systems operating in the single market and affected persons established in the EU. Article 2.1 specifies that the Regulation shall apply to economic operators placing on the EU market or putting AI systems into service. This forward-looking approach means that economic operators must first know whether the Regulation will apply to them and, if so, anticipate how the regulatory provisions will impact on their AI products not yet on the market, as well as on the services they plan to offer in the future.

In order to determine whether it has an impact on its scope of application, the provision establishes connection criteria that will imply that the Regulation applies to AI systems developed by providers or used by users established in third states. The connecting criterion will be that of a close link with the EU.

These connection criteria have consequences: firstly, legal uncertainty for operators who will find it difficult to determine whether they are subject to certain obligations; and secondly, the AIA may be applied to them unjustifiably because they do not have sufficient links with the EU[15] (for example, the application of European legislation to companies established in third countries in cases where they have a minimal link with the EU).

To achieve a comprehensive understanding of the prospective application of Article 2.1 of the AIA, it is essential to analyse it from several perspectives: that of proactive compliance; that of AI impact assessment and privacy-by-design; that of the risk-based approach; and that of AI ethics principles.

Let us look at each of these perspectives:

A) Proactive compliance.

Proactive compliance in the context of Article 2.1 AIA reflects the need for economic operators to address regulatory issues from the early stages of the design and development of AI systems. This approach not only ensures compliance with emerging regulations, but also aligns with ethical principles and societal expectations for responsible technology.

An essential component of a proactive compliance approach is internal training. Economic operators should invest in training programmes to ensure that their staff are aware of the requirements of the Regulation and understand how to apply AI system development practices in accordance with the

[15] *See ibid*, p. 6.

AIA. This is especially critical for those involved in the design and implementation of AI systems, as well as for compliance and risk management staff.

The testing phase of AI systems is a crucial moment for proactive compliance. Economic operators should implement rigorous testing procedures that not only assess technical functionality but also compliance with ethical and legal standards. This may require collaboration with external stakeholders, such as certification bodies or consumer groups, to validate the effectiveness and safety of AI systems.

Economic operators need to assess how their systems could be misused or how they could fail and the consequences for users and society at large.

B) Artificial Intelligence Impact Assessments and Design for Privacy.

A proactive compliance strategy implies that AI impact assessments must become an integral part of the AI product development lifecycle. These assessments should go beyond technical considerations and also address the social, ethical and legal implications of AI systems. The 'privacy by design' approach requires economic operators to integrate personal data protection measures from the product conceptualisation stage, ensuring that user privacy is not a secondary consideration but a central pillar of AI systems.

C) Risk-based approach.

The AIA categorises AI systems according to the level of risk they present. A proactive approach requires economic operators to identify the level of risk of their AI systems at the outset, i.e., prior to their introduction to the market, commissioning and/or use, by adapting their development processes to meet the necessary standards. For example, a high-risk AI system will require compliance with the following requirements: 1) establishment of a risk management system; 2) quality of the data sets used; 3) documentation and recording; 4) transparency and disclosure of information to users; 5) human oversight; and 6) robustness, accuracy and cybersecurity.

A key tool in this process is the AI impact assessment, which examines potential consequences before systems are deployed. These assessments should consider not only the intended use cases, but also hypothetical scenarios in which the system could be employed unintentionally. By anticipating these scenarios, operators can design more resilient systems.

Once risks have been identified, it is essential to develop mitigation strategies. This may include implementing technical safeguards (such as monitoring systems and early warnings), as well as creating policies and procedures that limit the use of AI to safe and ethical applications. Continuous training and skills upgrading of staff operating AI systems is also crucial for effective risk mitigation.

D) Ethic Principles in Artificial Intelligence.

Economic operators must also ensure that AI systems are in line with recognised ethical principles, such as transparency, fairness and non-discrimination. This implies a commitment to creating AI systems that are understandable to users and whose decisions can be justified. In addition, there must be safeguards to prevent bias and discrimination, which requires constant review and updating of AI models to reflect societal values.

Ethical and responsible AI design must consider the potential impact on individuals and society. AI systems must be designed in a way that respects human rights, dignity and democratic values. This refers to avoiding bias and discrimination and ensuring that automated decisions are fair and non-discriminatory.

III. Implementation of Article 2.1 by competent national authorities

The application of Article 2.1 of the AIA by the competent national authorities is "curious" to say the least; contrary to the GDPR, the AIA does not establish a right for natural or legal persons to lodge a complaint with national supervisory authorities for non-compliance with the provisions of the Regulation by providers, users or any other AI system operator. Moreover, its application implies the need to establish a framework for international competition between such authorities. Article 2.1. of the AIA sets out the material scope of the regulation, defining what is meant by AI systems and establishing the basis for their regulation, supervision and oversight. A number of factors are essential for the effective application of Article 2.1: the need for international cooperation; harmonisation of regulatory standards; and interaction with international law.

Let us look at each of these factors:

1. The need for international cooperation

International cooperation in the supervision of AI is a crucial component in the implementation of Article 2.1 of the AIA by competent national authorities. Given that AI knows no borders and can have significant impacts in multiple jurisdictions, the need for a coordinated approach is imperative. National competent authorities must therefore build bridges of collaboration and share information and resources to ensure effective supervision.

One of the biggest challenges for international cooperation is the diversity of regulatory frameworks. Each country may have its own approach to AI regulation, based on its cultural values, social norms and political priorities. This can lead to discrepancies in the interpretation and application of

AI regulations. Authorities should therefore work towards harmonising these approaches to enable a more homogenous regulatory ecosystem[16].

International cooperation does not stop at formal regulation; it also involves training and knowledge sharing. National authorities can greatly benefit from joint training programmes, staff exchanges that promote a common understanding of challenges and best practices in AI supervision.

Looking ahead, international cooperation in AI regulation is expected to strengthen further. With the rapid evolution of the technology, national authorities will need to remain proactive in their collaboration.

The implementation of Article 2.1 of the AIA by competent national authorities within the international context requires a continuous and concerted effort. By working together, authorities can ensure that AI regulation is effective, fair and non-discriminatory, protecting citizens and encouraging responsible innovation globally.

2. Harmonisation of regulatory standards

The harmonisation of regulatory standards for AI at the international level is a complex process, but essential to ensure that the technology is developed and applied safely and ethically in diverse socio-economic contexts. The implementation of Article 2.1 of the AIA by competent national authorities is directly influenced by the degree of consistency that regulatory standards can achieve at the global level.

AI regulatory standards cover a wide range of considerations, from security and privacy to fairness and transparency. The existence of numerous regulatory approaches reflects the diversity of values and objectives of societies around the world. However, this diversity can also result in a fragmented landscape that hinders international cooperation and trade[17].

A key component of harmonisation is the development of standardised certification and testing schemes. These schemes will allow authorities to assess and certify AI systems according to internationally agreed criteria. They thus facilitate mutual trust and recognition of the conformity of AI products and services.

To achieve effective harmonisation, it is crucial that different regulatory frameworks are interoperable. This means that the regulations of one juris-

[16] *See* Corcoy, M., "La inteligencia artificial en el derecho español", *Revista de Derecho y Genoma Humano*, n.º 54, (2021).

[17] *See* International Standards Organization (ISO), *ISO Standards for Artificial Intelligence* (2022).

diction should not conflict with those of another and that economic operators should be able to navigate easily between different regulatory systems without having to comply with contradictory requirements. One of the ways to achieve this, as will be discussed later, is through the use of multilateral and bilateral agreements between third states and the EU.

The major concern in the harmonisation of regulatory standards is the protection of fundamental rights. European legislation places particular emphasis on data protection and privacy, and any harmonisation effort must ensure that these rights are not compromised.

In practice, harmonisation of regulatory standards may involve setting up international working groups, drafting consensus documents and conducting comparative studies.

Harmonisation of regulatory standards is essential to create a safe and equitable global environment for the development and use of AI. Through the collective effort of competent national authorities, international cooperation, and the active participation of entities from all sectors, a regulatory framework can be achieved that not only protects individual rights, but also promotes innovation and economic growth.

3. Interaction with International Law

The interaction of the AI regulatory framework with international law is a rapidly evolving field, with multiple implications for trade, diplomacy, and global governance. The implementation of Article 2.1 of the AIA by national authorities must consider how local regulations align, complement or, in some cases, may conflict with existing international obligations.

One of the first considerations is how AI regulations fit into the fabric of previously established international treaties. For example, World Trade Organisation treaties include provisions that could be interpreted to address aspects of AI marketing and standards. EU regulations will need to respect these existing agreements or seek their modification when they relate to AI.

In addition, international humanitarian law and human rights law set limits on the development and use of technologies that may be employed in military contexts or that may affect individual rights. AI regulations should ensure that AI systems are consistent with these principles, prohibiting uses that violate international law[18].

The EU, with its progressive approach to AI regulation, has the oppor-

[18] Spanish Ministry of Foreign Affairs, European Union and Cooperation, Regulatory Diplomacy and AI, (2023).

tunity to lead in the formulation of international legislation in this area. The policies and regulations it develops can serve as a model for future international treaties and legislation in other countries, promoting high standards of data protection, privacy, and security.

IV. Application of Article 2.1. by the Courts of Justice

AIA also raises issues of jurisdiction and extraterritorial application. The EU must work to ensure that its regulations are respected beyond its borders, which is a significant challenge in the globalised digital space. This may require bilateral or multilateral agreements, as well as constant dialogue with other jurisdictions to ensure cooperation in the supervision and enforcement of these regulations.

The Council of Europe and other international human rights organisations are critical forums for dialogue on how AI applications may affect human rights. EU regulations can influence the creation of global guidelines to ensure that AI is developed in a way that respects human dignity and fundamental rights.

It is essential that EU regulations consider the impact on developing countries, which may lack the infrastructure to meet stringent standards. International development cooperation and technical assistance will be crucial to ensure that AI is a tool for advancement and not a source of division.

AIA has the potential to shape not only the European regulatory landscape but also the global governance of AI. However, to be effective and fair it must be articulated within the framework of international law, respecting existing treaties and contributing to the development of new standards and principles. This requires a concerted effort for international cooperation, technology diplomacy and the promotion of an inclusive and holistic approach that embraces all regions and sectors of society.

The AIA by the EU introduces significant challenges in terms of jurisdiction and territorial scope. The inherently global nature of AI and its associated industry demands detailed scrutiny of how regulation in one territory may influence (or be implemented by) internationally operating entities. The application of Article 2.1 of the AIA makes clear the need for a holistic and globalised approach to regulation, with an emphasis on international cooperation and regulatory harmonisation.

The overriding concern regarding jurisdiction is the extraterritorial scope of the AIA. That is, the EU must define how its rules will affect companies and entities outside its territory that produce or provide AI systems used within the EU.

The implementation of Article 2.1 of the AIA by national competent authorities emphasises the need for global dialogue and collaboration to develop a harmonised and equitable approach to AI regulation. Ultimately, the EU's success in regulating AI will not only be measured by the effectiveness of its domestic legislation, but also by its ability to influence and be part of a cohesive global regulatory framework.

National competent authorities are not mere enforcers of EU AI legislation; they are active participants in the global regulatory landscape. By implementing Article 2.1 of the AIA, these entities contribute to the formation of an international landscape that is more cohesive, fair and balanced. Their role goes beyond policy implementation, extending to influencing the governance of AI globally, which is crucial to address the challenges that the technology presents in an interconnected society.

Article 2.1 of the AIA itself also notes that the application of the Regulation may sometimes arise in the context of a civil action brought before a court concerning, for example, extracontractual civil liability arising from the malfunctioning of an AI system, or a breach of a contract between a provider of AI systems and a user, or a dispute between any such provider and an individual who is a party to a contract for the provision of services using such systems. In such disputes, non-compliance with the requirements or obligations set out in the Regulation for the different categories of AI systems may be invoked as a basis for the claim.

In civil or commercial disputes, the international jurisdiction of the court of the Member State before which the action is brought will be determined by the "Brussels I bis" Regulation -the courts of the state where the alleged injured party has his habitual residence, those of the place of work or where the breach of the AIA took place having jurisdiction- and the applicable law by the "Rome II" Regulation -if the dispute is about non-contractual civil liability- or the "Rome I" Regulation -if the dispute is about the breach of an international contract. However, the substantive applicable law (the *lex causae*) will be the AIA itself, and the foreign law of a third State may not be applied.

V. Extraterritorial application of the European Artificial Intelligence Act

The aim of this broad territorial scope is that the protection offered by the GDPR 'travels' with personal data wherever it goes in a globalised society where data crosses borders at the click of a button. The EU is guided by the reasoning that providing protection only for data processing that takes place within European borders would not be sufficient. This measure also seeks to provide a level playing field for European companies without creating stricter regulation that would place burdens on them alone. The extra-territorial application of the GDPR means that any company wishing to access the European market to offer its services and goods and to process 'European' personal data must comply with these rules even if its head office is in a third country[19].

Extraterritorial application of legislation is not new[20], but it can be seen to be gaining a lot of traction in aspects of Internet regulation[21].

The GDPR has been heavily criticised, as with the number of companies falling under the criteria worldwide it is easier for multinationals to adapt to the GDPR while it is very costly for *SMEs*[22] In addition, data protection authorities in Member States have limited resources, so as there will be more foreign companies that do not comply with the GDPR than there are resources to investigate them, the actual implementation of the GDPR will necessarily be arbitrary, undermining the legitimacy of any enforcement action taken[23]. However, such extraterritorial enforcement could be considered legitimate and argues that the EU is equipped with the relevant tools to enforce the GDPR abroad, even if they need to be further developed[24]. This approach, although not without drawbacks and challenges for state

[19] *See* Gascón Marcén, A., "The extraterritorial application of European Union Data Protection Law", *Spanish Yearbook of International Law*, n.º 23 (2019), pp. 413-425, p. 415.

[20] *See* Dover, R. and Frosini, J., *The Extraterritorial Effects of Legislation and Policies in the EU and US,* European Union, Brussels 2012. According to Gallego, although the EU has never been a complete advocate of extraterritoriality, it is beginning to redouble its exercise through territorial extension, which makes it possible to control conduct that, although carried out abroad, has an impact on the general interests of the EU. See Gallego Hernández, A. C., "La aplicación de la extensión territorial del Derecho de la Unión Europea", *Cuadernos Europeos de Deusto,* n. º 63 (September) (2020), pp. 297-313, available at: https://doi.org/10.18543/ced-63-2020pp297-313.

[21] *See* Internet Society, *The Internet and extra-territorial effects of laws,* Internet Society, 2018, p. 1. The Internet Society warns that many states are imposing rules that spill over to the internet elsewhere, hinder innovation, deter investment in their own countries, and risk creating new digital divides that harm their own citizens.

[22] *See* Scott, M; Cerulus, L; and Kaya LI, L., "Six months in, Europe's privacy revolution favours Google, Facebook", *Politico.eu*, 23 November 2018.

[23] *See* Svantesson, D. J. B., "European Union Claims of Jurisdiction over the Internet – an Analysis of Three Recent Key Developments", *Journal of Intellectual Property, Information Technology and E-Commerce Law,* vol. 9, no. 2 (2018), pp. 113-125, p. 118.

[24] *See* Azzi, A., "The Challenges Faced by the Extraterritorial Scope of the General Data Protection Regulation", *Journal of Intellectual Property, Information Technology and E-Commerce Law*, vol. 9, no. 2 (2018), pp. 126-137, p. 137.

interests and individual rights, solves one of the biggest problems European data protection law faced until then, which was the lack of jurisdiction over data controllers in third countries processing a considerable amount of Europeans' data[25].

European legislators were well aware that extraterritorial application of laws could have undesirable impacts. The GDPR itself in Recital 115 states that the extraterritorial application of some laws, regulations and other legal acts may be contrary to international law and may impede the protection of natural persons guaranteed in the EU under the GDPR; and therefore, data transfers should only be made in compliance with the conditions of the GDPR. Thus, we see that the GDPR provides for its own extraterritorial application, but excludes that of foreign laws in many cases. Such a conflict may arise, for example, when US authorities require data in the framework of a criminal investigation from a company located in their territory, but which relate to an EU resident contrary to the provisions of the GDPR, so that the company may be faced with conflicting legal obligations.

The problems are manifold and critics have good reason to be concerned, but the difficulty of ensuring the implementation of the GDPR or the lack of resources to do so cannot cause us to aim for lower standards of protection of fundamental rights; especially given how the GDPR has served to raise this level of protection not only in Europe.

To understand the nature of extraterritoriality in the AIA, it is essential to analyse the two key elements underlying its application.

The first element is the "offer" and "use" criterion. According to the Regulation, the regulations will apply not only to entities offering AI services in the EU, but also to those whose AI systems are used in the EU, regardless of whether that entity is established in the EU or not.

The second key element is the "effect" principle. The effect principle implies that, if an AI system has a significant impact on individuals or entities in the EU, then the law will apply. This extends even to systems developed and operated entirely outside the EU, highlighting the intention of the Regulation to protect its citizens from potential risks regardless of the location of the AI company.

Extraterritoriality in the AIA is also reflected in the obligations of non-EU entities. These firms must appoint a legal representative in the EU to ensure that they comply with the law and act as a point of con-

[25] *See* De Hert, P. and Czerniawski, M., "Expanding the European data protection scope beyond territory: Article 3 of the General Data. Protection Regulation in its wider context", *International Data Privacy Law*, vol. 6, no. 3 (2016), pp. 230-243, p. 230.

tact with the regulatory authorities. This is similar to the requirements set out by the GDPR and is critical to ensure that non-EU entities can be subject to oversight and sanctions if they fail to comply with the standards set.

This approach has significant implications for the global governance of AI: a) on the one hand, it sets a high standard that could inspire other jurisdictions to follow suit, promoting a form of "regulatory diplomacy"; and, b) on the other hand, it also raises questions about sovereignty and the balance of power in the regulation of emerging technologies.

Extraterritoriality, however, is not without its critics and concerns. Some argue that it could lead to conflicts of laws, where companies find themselves caught between incompatible regulations from different jurisdictions. Undoubtedly, the administrative and financial burden of complying with multiple regulatory systems can be onerous, especially for *startups* and *SMEs*. To address these concerns, the EU may need to collaborate with international partners to develop common standards or mutual recognition mechanisms to facilitate cross-border compliance. In addition, the EU must consider the economic impacts of its extraterritorial regulations and balance consumer protection with an enabling environment for innovation and trade.

Article 2.1(a) of the AIA is a connecting criterion informed by the jurisprudential doctrine of "targeted activities", used for example in the area of consumer contracts concluded on the Internet, or online infringement of unitary industrial property rights. This criterion ensures that the Regulation itself is applicable in situations with a close link to the EU.

Article 2.1(b) of the AIA is open to criticism for two reasons: a) to begin with, the use of the term "located in the EU" gives the Regulation itself an extremely broad scope of application. The application is unjustified as the situation has very little connection with the EU. This problem would be solved by amending the provision to limit its application to users established or habitually resident in the EU; and,

Indeed, the AIA does not apply to providers established in the EU who market their AI systems exclusively in third States, but it does apply to users established in the EU who provide their services in third States. The difference in treatment is unjustified. In both cases the link to the EU is the same. If the intention is that European users of AI systems should comply with the standards laid down in the Regulation irrespective of the country in which they offer their services, then EU-based providers marketing AI systems in third countries should also comply with those standards.

Alternatively, a case could be made for amending Article 2.1(b) so that the Regulation would only apply to deployers of AI systems where the output information generated by the system is used in the EU, regardless of whether they have their habitual residence or establishment on European territory or not.

The criterion of Article 2.1(c) of the AIA is a criterion that may lead to an unjustified extraterritorial application of the Regulation itself; and may be applicable in situations that are difficult to foresee for AI system providers established in third States[26] .

VI. Final reflection

The AIA is configured as a new global regulatory standard for AI, thanks to the extremely broad territorial scope of application granted to it in Article 2.1. This extremely broad territorial application of the AIA is not always justified, as the criteria for connection with the EU, as we have already pointed out, are sometimes "scarce".

Perhaps the use of the conventional, bilateral or multilateral route to extend European regulatory standards beyond our borders would be the most optimal way to help ensure compliance with AIA by providers and users based in third states; particularly in line with the European foreign policy tradition of consensus-building through bilateral or multilateral negotiations. International cooperation in the context of AIA is a crucial step towards creating a safe and ethical global environment for the development and application of AI. As the technology advances and its impact becomes globalised, working together becomes indispensable to manage its challenges and maximise its benefits.

In short, the extraterritorial application of the AIA leads us to reflect, furthermore, on whether we should abandon STEINER's "idea of Europe": Europe has always believed (and will always believe) that it will perish; that it can consolidate and progress; and, ultimately, serve as a mirror and competition for other countries.

Hopefully, the extraterritoriality of the AIA (and its character as a "containment and reorientation rule") does not turn the EU into a lagging island in the world that does not allow us to make progress in innovation. However, we will have to wait a few more years to analyse the true extraterritorial impact of AIA.

[26] *See* López-Tarruella Martínez, A., "El reglamento... "cit. pp. 14-17.

THE EXCLUSION OF NATIONAL SECURITY, DEFENCE, AND MILITARY ARTIFICIAL INTELLIGENCE SYSTEMS FROM THE REGULATION AND THE APPLICABLE LAW

Ángel Gómez de Ágreda

Lecturer at the Universidad Politécnica de Madrid. Spanish Ministry of Defence. Odiseia[1]

I. Introduction

Artificial Intelligence (AI), as a concept, is a very broad field to consider its regulation as a whole. The approach being followed by all countries is to study the different uses it has. This is still a difficult task due to the wide casuistry of each one of them, the constant evolution to which these technologies are subject and, not least, the substantial advantage that their use offers to those who are at the forefront of their study.

The latter is particularly true when it comes to military uses – in a broad sense. As has historically been the case with all technological advances, those who make the most intensive use of them tend to resist their restriction, while the less advanced ones invoke all sorts of risks and threats. The case of the banning of crossbows and longbows at the Second Lateran Council on the grounds of the indignity of killing (nobles and knights) at a distance has remarkable parallels with the present day.

The ability of many technologies – most notably AI – to be used for both beneficial civilian applications and others of a warlike and destructive nature is known as the dual use of technology. This dual nature forces a much broader view of the possible applications and, consequently, of the aspects to be regulated.

At the same time, the mere fact of including the warlike uses of these media is likely to generate a certain social rejection of their development or, at least, condition it to greater scrutiny. In a context of very rapid advances, the great power accumulated by technological corporations and that of their benchmark states combines, in this case, to slow down any regulatory process that might offer a competitive advantage to a rival.

For all these reasons, and because of the complexity and specificity of

[1] ORCID: 0000-0003-1036-6324. This work is carried out in the framework of the Project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by MCIN/AEI/10.13039/501100011033/ERDF, EU.

the subject, most legislation and ethical codes relating to AI leave aside the treatment of its use in warfare.

In many cases, they directly ignore these possible uses in their development. In others, they deny the legitimacy of this use and advocate its prohibition with little or no argumentation or semblance of realism. Finally, many other codes recognise the possibility of this dual use, but decline to include military aspects in their treatment. In this way, they consciously distort – and whitewash – the image of AI.

If this is the case in "civilian" (non-military) codes, the codes that have recently been developed to deal specifically with this use also tend to suffer from two common problems. On the one hand, although many include in their articles the need not to do so, they condition the beneficial development of these technologies. On the other hand, they tend to associate military uses of AI with lethal autonomous weapon systems (LAWS), i.e., "killer robots".

In reality, most AI applications in the military field are not associated with driving autonomous platforms or vehicles, but rather with data analysis, decision-making or logistical tasks. Despite their often non-lethal nature, the specificity of the framework in which they are developed (and the legislation applicable to it) and the effects they can have in warfare also require specific treatment.

Indeed, it has been argued that the ethical principles and legal rules that are developed for the military domain could have important lessons and conclusions for the regulation of civilian systems. The lethality associated with many of these applications and the visibility of the effects they cause illustrate factors common to any AI-enabled system better than other uses.

In this regard, it is important to note that a distinction needs to be made between the design, development and commissioning of AI-enabled systems, and the use that is made of these systems. For example, technology-specific regulation should focus on ensuring that biometric data processing systems are free of racial, gender, class or other biases, and that they take into account the need to protect the privacy of individuals. Meanwhile, other codes should consider whether this application is used for the selection of personnel for a company, for court support in relation to a list of suspects, or for the selection of targets for an autonomous perimeter defence system in a military installation.

Pretending to ignore possible warfare applications of technological developments is as dangerous as forcing the renunciation of these same developments on the basis of their potentially harmful use.

The same can be said of the necessary differentiation between the regulation of technology and the regulation of the techniques of its use. One of the main advances introduced by the study of AI-enabled systems is a greater understanding of human cognitive mechanisms and a greater ability to use algorithms to affect them. In addition, the reduced adaptability of mechanical systems means that the environment is being redesigned to help machines understand it, giving Artificial Intelligences a competitive advantage over natural ones.

Many ethical codes – and practically all companies and governments – emphasise the need to avoid the opportunity cost of delaying AI research and development in general and for the benefit of humanity in terms of the negative consequences of its misuse. It is therefore urgent to put in place the necessary controls to minimise the perverse effects of the dual use of these technologies as soon as they are being designed.

The Council of Europe's Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law follows the same line of excluding national defence from the scope of its competences. In this respect, it barely qualifies the AIA in a reminder that the law, or rights, as the case may be, remains applicable to these systems as long as it is not explicitly excluded. This is a vague reference that has been used previously in other cases such as cybersecurity (Tallinn Manual) or privacy.

This paper will discuss the treatment of military uses of AI in texts relating to the first aspect, the technological one. In addition, a review will be made of the international legislation applicable to AI-equipped systems and their use *ad bellum* and *in bello*. In the latter case, the references are generally indirect and general, not exclusively applicable to the case of AI.

II. Ethical codes for general-purpose Artificial Intelligence

In the first AI codes of ethics (2016 with Satya Nadella (Nadella, 2016) to 2019) the tendency is to ignore possible war or dual uses of AI. In some cases, it is explicitly excluded, in others it is simply not mentioned or advocated to be banned. The exceptions are the IEEE Ethically Aligned Design V.2 (IEEE, 2016) and COMEST (the World Commission on the Ethics of Scientific Knowledge of UNESCO) (COMEST, 2017). Both advocate strengthening human control and responsibility over machines, and the development of specific ethical and legal codes for such autonomous systems.

In almost all cases, the use of AI in warfare is wrongly associated with lethal autonomous systems, when most of it – and the most dangerous – has to do with the decision-making process, and with logistical and intelligence support. Its use in cyber-attacks, while not normally lethal, is also very dangerous.

At (Gómez-de-Ágreda, 2020) these initial ethical codes relatingo to AI are summarized, and their relationship with its military applications is detailed. Subsequently, numerous codes have been developed specifically for use in warfare. Unlike the first ones – logically – the latter are drafted by the ministries of defence of the various governments and therefore suffer from a lack of pluralistic vision.

In the study of the different codes, it is therefore necessary to keep in mind who the author is and what his or her interests are. While business tends to justify the need for continued development and innovation, civil society tends to be very wary of the harmful uses of AI. However, the caution is more related to its possible perverse "civil" uses than to the duality in the usability of algorithms.

III. Dual use of technologies

The ethics of AI have often been related to those of other technologies, such as nuclear. In both cases, there are numerous and very important civilian applications that more than justify research in this field. They also share the criticality of the possible harmful uses of their military versions: a nuclear attack or the possible consequences of general AI.

However, while both are present in the popular imagination, AI is still linked to the realm of fiction, hypothesis and dystopia. There is, in reality, no real perception of danger from the use of algorithms. The dominant narrative focuses on the risks arising from future developments in the medium to long term, while leaving aside – in many cases self-interestedly – the real threats of the present uses of this technology. In this way, rather than concern, a certain morbidity is created, detached from reality, which distances the general public from the urgency of regulating its use.

Other technologies associated with the military – such as those used in electronic warfare – are too far removed from the everyday life of the population to generate social alarm. At the same time, the lack of an efficient market in the civilian sector severely limits the duality of their use and the need for civilian control.

The intangibility of AI, often distributed as open source with little or no oversight over which applications it is integrated into, contributes to its invisibility.

Ethics and legislation concerning the use of AI should, on the contrary, be associated with other technologies more closely linked to freedom and human will. These include neuroscience and biotechnology. In both cases, these are disciplines that affect the very nature of the human being, something that AI can do indirectly. As mentioned above, the programming of algorithms has benefited greatly from advances in the understanding of the human cognitive side, and vice versa.

Within this parallelism, one could apply concepts such as the "Dual use research of concern", used in The Human Brain Project (Aicardi, 2018)which applies to the initial phases of the development of these technologies. The current trend is, moreover, to make joint use of all these disciplines to model the individual in all its facets, from genetics to cognition.

In any case, none of these disciplines has the "democratic" character of AI that makes its evolution widely distributed among numerous actors, many of them not even professionals in the field. AI's capacity for uncontrolled growth has no parallel in other modern sciences.

Even autonomous systems used in combat are not rejected outright by public opinion, which adapts as operational needs and the official narrative explain the advantages – undeniable, moreover – of their use.

This lack of awareness of the risks associated with the dual use of AI is related to the fact that systems are interpreted as a whole, without disaggregating the various AI-enabled components from each other and from the platforms that host them. An autonomous combat drone shares many common characteristics with an unmanned car. But a further breakdown allows us to identify subsystems that, on their own, do not appear to pose a threat to humans: image identification systems, for example.

The duality of AI use is therefore not only limited to whole systems, but to the different algorithms that allow them to function and which, applied on different platforms or in conjunction with other algorithms, can generate different threats (in the case of image identification, to continue with the example, in relation to privacy).

IV. The CCW Decalogue of Ethical Principles for Lethal Autonomous Weapon Systems

The United Nations Convention on Certain Conventional Weapons (which can be particularly harmful) has been meeting every six months since 2013 to try to reach international agreements to regulate this type of weapon. This type of forum provides great legitimacy as states and civil society are represented (ICRC, for example), but lacks coercive capacity and clear leadership. This is evident in the desiderative rather than impositive nature of their statements.

In fact, the major nations and industries involved are the first to artificially delay any process of adopting legislation that might constrain their use as long as they have the upper hand.

The CCW has only managed to produce a decalogue (in 2018) in which it affirms the applicability of international humanitarian law to any type of war regardless of the weaponry used.

At its 2018 meeting, Austria, Brazil and Chile submitted a proposal that culminated in the launch of an open working group to negotiate a binding agreement (Austria et al., 2018). Simultaneously, the CCW reached an agreement – in this case, not legally binding – on a set of ten principles (CCW, 2018b) which, while not exclusively ethical in nature, represent the greatest progress of the working group so far and still represent a starting point for further expansion:

1. International humanitarian law remains fully applicable to all weapons systems, including the potential development and use of lethal autonomous weapons systems.

2. Human responsibility for decisions related to the use of weapon systems has to be maintained as legal responsibility cannot be transferred to machines. This principle should apply throughout the entire life cycle of the weapon system.

3. Legal responsibility for the development, deployment and use of any weapon system emerging within the CCW framework must be ensured in accordance with applicable international law, including for the operation of such a weapon system within a responsible human chain of command and control.

4. In accordance with state obligations under international law, in considering the development, acquisition or adoption of a new weapon, means or method of warfare, it must be determined whether its use would, in some or all circumstances, be prohibited by international law (applicability of the Martens Clause) (Hague Convention (II) on the Laws and Customs of War on Land, 1899; Ticehurst, 1997).

The Martens clause, mentioned repeatedly throughout the CCW discussions, establishes the need to apply to novel types of tactics or weaponry the same criteria that, as a matter of common sense, can be extrapolated from the generally applicable standard. This diffuse logic of common sense is even further beyond the reach – at least in the current state of the art – of autonomous systems and underlies much of the public opinion hostile to these systems (M. C. Horowitz, 2016). Similarly, the principle of humanity – and displays of compassion – cannot be expected to be respected by a system designed to optimise combat advantage.

5. Physical and non-physical safeguards (including cybersecurity against hacking or data impersonation), the risk of acquisition by terrorist groups and the risk of proliferation should be considered during the development or acquisition of new weapon systems based on emerging technologies in the LAWS area.

6. Risk assessment and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapon system.

7. Consideration should be given to the use of emerging technologies in the area of LAWS in ensuring compliance with international humanitarian law norms and other international legal obligations.

8. In the development of potential policy measures, emerging technologies in the area of LAWS should not be anthropomorphised.

9. Discussions and any potential regulatory measures taken in the context of LAWS should not prevent the advancement of or access to peaceful uses of smart autonomous technologies.

10. The CCW (...) seeks to strike a balance between military necessity and humanitarian considerations.

It is interesting to note how the Decalogue rightly refers to aspects linked to the use of technologies, rather than to the technologies themselves. Since its drafting, however, international political polarisation has prevented further progress or even the practical implementation of the precepts it contains.

Most general ethical codes include beneficence as their first principle. In the case of weapons systems, relative beneficence is often referred to, although this is also a highly debatable concept. In support of this "relativity" – compared to an equivalent action carried out by a human intelligence – it is argued that artificial sensors and algorithms are more discriminating than human senses and reasoning.

Thus, the application of, for example, the principle of distinction (between combatants and non-combatants) would benefit from the increased sharpness of cameras and microphones. Many non-governmental organizations and academics question this argument and the actual ability to discern between aggressive or peaceful attitudes. This lack of ability to go beyond the directly measurable is a further argument for keeping alive the periodic review of the ethical criteria applicable to the use of AI-enabled systems.

V. Significant differences between Civilian and Military uses of Artificial Intelligence

The degree of complexity of the military environment -and especially of the war environment- is much higher than that of the civilian one. Firstly, because the scenario in which it has to develop its action is much broader in the case of the military (such as, in the case of unmanned vehicles, the need to navigate through unprepared terrain and not only on conventional roads). In addition, military systems have to deal with potential adversarial action, rather than collaborative action with other systems, as is the case in the civilian world.

Clearly, military systems will often be linked to critical decisions and actions, which is only rarely the case in the non-war social environment. This concerns not only autonomous weapon systems with lethal capabilities, but also decision-making or coordination of military operations. In many of these cases, the conclusion concerns human lives and therefore involves ethical criteria that may not be as relevant in other jobs.

This criticality forces a peculiar interpretation of the military use of AI-enabled systems. A specific case can be found in the search for predictability of algorithmic results which appears as a requirement to be met in many ethical codes. However, a predictable system is extremely vulnerable in an adversarial environment. If the enemy can foresee the system's reaction, it can also counteract it.

As with 'beneficence' – another of the most widespread ethical principles – which is transformed into 'relative beneficence' in the military domain, predictability is transformed into mere 'reliability' (ICRC, 2018). (ICRC, 2018). In other words, consistency in the achievement of objectives is demanded, but automatic repetition of the means to achieve them is rejected. Systems must be predictable only to the user, but opaque to the adversary.

Of course, in the field of law, military activity is affected by specific differentiated legislation and is treated very differently from civilian or criminal law. The application of the rules of International Humanitarian Law, the specific conventions (Hague and Geneva) and the Law of War (Law of Armed Conflict) are specific to the military and war environment. Lethality, which is banned in the civilian sphere, becomes a starting point in the military sphere. It is not its exclusion that is sought, but its restriction to specific circumstances and to specific purposes.

An additional risk is posed by the use of AI-enabled systems that are not yet mature in their development. This is more likely in the military than in the civilian domain because of greater risk tolerance and less legal oversight prior to use. Historically, this greater flexibility has led to numerous innovations that have subsequently been transferred to the civilian world. However, these have often come at a painful price and in serious violations of human dignity as understood at the time.

VI. Applicability of International Law to artificial intelligent systems

The premise is therefore established that International Law, in particular International Humanitarian Law, has absolute and indisputable application in the use of AI systems. The consensus reached in the CCW leaves no room for doubt in this regard.

There is no disagreement with the CCW's criterion since the preamble to the Hague Convention (II) of 1899 relating to the Laws and Customs of War on Land already incorporates the so-called "Martens Clause" which states: "Pending the publication of a more complete code of the laws of war, the High Contracting Parties deem it appropriate to declare that in cases not covered by the Regulations adopted by them, the populations and belligerents remain under the protection and rule of the principles of International Law as they result from the established usages among civilised nations, from the laws of humanity and from the requirements of public conscience."

For those tempted to argue that, since then, more comprehensive codes have been published than those claimed by the wording of the clause, it should be recalled that both the 1949 Geneva Conventions and the two Additional Protocols of 1977 reaffirm and emphasise it, even including it in their Preamble.

Article 36 of the Geneva Convention itself contains an additional and complementary argument to the Martens Clause. The wording of this article demands that, "in the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party has the obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party".

The drafting of specific rules for AI-enabled systems is, however, generating a major struggle between the major powers, industry and those countries with less access to these technologies. Always on the premise of the universality of established law, interpretations continue to diverge considerably depending on the interests of one or the other.

VII. Accountability and meaningful human control

Responsibility in AI systems is often very difficult to establish as several systems combine their actions in a complex task. Information gathering, information processing and decision making (and execution) may be assigned to different systems at different times.

The concept of shared responsibility appears now. Here, the responsibility continues to lie with the humans behind each process, but also behind each part of the chain of creation of devices and systems. In this way, the designer retains his share of responsibility for an unlawful use of his work. The failure to create safeguards to avoid this scenario falls on him. The same applies to the developer of the design, the integrator, the distributor and, of course, the end operator of the system and the whole chain of command that contributes to its use.

It is, in any case, always the human – as a species – who bears the responsibility for the acts committed by machines. Just as in aircraft accident investigations, even mechanical failures can be attributed to human actions or omissions in design, development, operation, training, maintenance and so on.

It is therefore not possible to rely on the beneficial intentionality of a design, but rather to envisage a possible use that is not. AI is not merely a tool for altering an environment, but is potentially an environment in itself and, as a consequence, places greater responsibility on its creators.

Dispute with the concept of meaningful human control. Definitions from different organizations.

CNAS[2]		Article 36	ICRAC[3]	ICRC[4]
Human participation	Informed conscious decisions	Judgement and timely human action	Cognitive participation. Perception and action	Human intervention at all stages
Required information	Sufficient information on weaponry, purpose and context	Accurate information on the technology, the objective and the context	Nature of the target and collateral damage. Full situational and context awareness	Information on the weapons system and context

[2] Center for New American Security, a US think tank.

[3] International Committee for Robot Arms Control.

[4] International Committee of the Red Cross.

Armament design	Proven weaponry. Trained human	Predictable, reliable and transparent technology	Suspension or abortion of the attack	Predictability and reliability
Legal requirements	Sufficient information to ensure legality	Accountability to some extent	Need for the attack to be appropriate. Compliance with IHL	Accountability and compliance with IHL

1 Table. Differences in basic concepts by different agencies. The author in (Gómez-de-Ágreda, 2020)

Responsibility lies not so much with the final executor of the action as with the one who makes the decision to carry it out. Human autonomy, the Anglo-Saxon concept of agency, has to do with this capacity to decide and must be disassociated from the physical act of "pulling the trigger". In fact, the attribution of responsibility to the executor can result in an unfair discharge of responsibility. The operator becomes the "scapegoat" for a decision that he has not taken, or whose adoption is vitiated by the biases introduced by the algorithms that have obtained, selected and interpreted information that has already been digested.

Degrees of autonomy cannot therefore be defined in terms of the proximity of the point of human intervention to the final decision. In many models, for example, the execution is always carried out by the machine, but the decision is made by a human with varying degrees of freedom.

Nor is lethality a factor to be taken into account in the development of ethical and legal codes for the use of AI-enabled systems in the military domain. Lethality is an inherent factor in the act of warfare and is the starting point for the legislation that regulates it.

In the case at hand, most of the systems equipped with AI for use in the Armed Forces are not directly related to the use of force and, much less, are lethal or form part of a weapon. However, following the above reasoning, it is not appropriate to differentiate between the principles applicable to one or the other.

In a very special way, it must be borne in mind that much of modern conflict – and life in general – is fought in the virtual and cognitive realm. Regardless of their direct effect on the material world, cyber and disinformation acts are clear examples of aggressive actions being employed as part of military operations. The tools used in them should, therefore, be considered in the same way as weaponry that directly produces death or destruction.

VIII. Use of Artificial Intelligence in Security and Defence under the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law

Article 3 of the first chapter of the Convention sets out the scope of the Convention. The fourth point of the article, the most concise of all, simply excludes matters related to national defence from the scope of the document.

Previously, the second point of the article also exempts the parties from applying the content of the Convention to activities related to their national security. The only qualification it contains is that these activities are understood to be carried out with respect for international law and human rights, as well as democratic institutions and processes; a sort of Martens Clause applied to Artificial Intelligence.

While the latter is somewhat more restrictive than the total exclusion from the application of the convention that is envisaged for defence-related cases, the ethical and normative vagueness itself, and the different interpretations discussed in this chapter, provide little anchorage for such restrictions.

It is significant that a similar reference to the applicability of the law to any warfare system, regardless of whether it is digital or based on Artificial Intelligence, is not included in the wording of the defence section.

The AIA already excludes security and defence from its scope of application. However, some nuances introduced by the legislator can be seen in its wording.

The fact that the regulation does not affect state competences in matters of national security extends to any entity entrusted with these tasks. This leaves the door open to the possibility of outsourcing tasks related to national security in companies or entities outside the administration in a sort of counterpoint to the extension of state responsibility to other entities when they act on its behalf.

This aspect is far from minor or inconsequential as it reflects the increasing involvement of contractors and corporations in national security-related tasks, especially in technological or technology-supported tasks. Extending the umbrella to these cases could lead to abuses or biased interpretations of its spirit.

Secondly, it is noteworthy that the exemptions only apply to systems that are used exclusively for military, defence, or national security purposes. The overtly dual nature of digital technological developments, and in particular those linked to Artificial Intelligence, raises questions as to whether there are designs whose use is closed to activities other than those originally envisaged. No provision is made for this in legislation, although it is covered by various codes of conduct or codes of ethics mentioned in this chapter.

IX. Conclusions

Although the international community does not question the applicability of International Humanitarian Law to the use of Artificial Intelligence weapon systems, it is not foreseeable that an executive consensus on their use can be reached in the coming years.

For the time being, the United Nation CCW has given itself until the end of 2025 to study the issue further in order to develop ethical and legal criteria on this issue (Lipton, 2023). (Lipton, 2023). It is precisely the most advanced countries that are most reluctant to regulate the activity of media from which they themselves derive the greatest benefit.

The US has made its position clear that it will not accept impositions while emphasising the importance of meaningful human control over the actions of these systems for the time being. Russia disagrees that this is a priority. Most countries believe that a delay in legislation – which a lack of consensus and will leads to – will result in a massive presence of these systems. (Hicks, 2023) and a lack of control over the degree of autonomy they may be endowed with.

Washington seems to have opted for the route of facts by publishing internal policies, both from the Department of Defence (*Autonomy in Weapon Systems*, 2023) and the State Department (US Department of State, 2023) and inviting other countries to adopt them as written.

With the precedents for comparable weapons, it is to be expected that no significant breakthrough will occur until the systems have been used in a major power conflict or until the technology is sufficiently mature to be considered non-disruptive. Liabilities will be enforceable on the basis of existing legislation, including Article 36 and the Martens Clause as the most significant underpinnings.

Meanwhile, the potential dual use of these technologies – not only complete systems, but also individual components – remains equally unregulated in the civilian sphere.

The urgency of regulating lethal autonomous weapons systems does not stem solely from their own use, but also from the ability of the ethical and legal principles adopted to serve as a reference for other uses of AI.

Both the AIA and the Council of Europe Convention, however, exclude the military – and, to a large extent, national security – from their competences. On the one hand, the current geopolitical situation and, on the other, the interests of industry make a more precise approach unfeasible at this time. In the first scenario, the challenges of restricting arms development in a pre-war context like the current one, along with the limited opportunities for consensus among the competing parties, make it impossible. In the case of industry, the incipient nature of developments and their dizzying pace favour, in any case, the promotion of research and innovation at all costs rather than the adoption of restrictive measures.

It seems unlikely that legislation will impose restrictions on Artificial Intelligence developments in general because of their economic and strategic value. Much less that, at this stage, any measures restricting the possibility of gaining military advantage from the application of these technologies could be envisaged.

THE ARTIFICIAL INTELLIGENCE ACT
AND THE GENERAL DATA PROTECTION REGULATION

Jesús Jiménez López
Director of the Council for Transparency and Data Protection of Andalucia

I. Introduction

While it is true that the fundamental right to the protection of personal data and its regulatory and institutional guarantee structure must be applied regardless of the technological tool used in its processing, it is also true that Artificial Intelligence implies specific challenges in this area. The purpose of these lines is to check to what extent the AIA responds to it, in the sense of enabling, assisting, or hindering the application of the GDPR to these systems. Therefore, this document aims to provide a brief analysis of the relationship between the AIA and the General Data Protection Regulation.

A brief comparison of both legal texts, specifically the provisions relating to their purpose, allows us to verify that the GDPR refers in all cases to the processing of personal data, whatever the form or technology by means of which this takes place, either because it is necessary for the development and use of Artificial Intelligence systems (AIS), or because such systems are the means to carry out the processing, as a specific task.

For that reason, the first clarification to be made is that it is not intended to make a study on the application of the GDPR to Artificial Intelligence systems, in consideration of their peculiarities, during their life *cycle,* or in the context of their *value chain*, a work advanced by authoritative doctrine and independent supervisory authorities[1].

In a way, we want to contrast two regulatory texts, GDPR and AIA. To what extent they relate to each other, overlap, complement or modify each other, whether or not in a clear way, always from the point of view of the legal certainty necessary for the preservation of the right to the protection of personal data, in the context of the essential technological development. In short, it is a question of outlining a framework of certainty in the application

[1] Palma Ortigosa, Adrián (2022); AEPD. (2020). Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial; ICO UK. (2023). ARTIFICIAL INTELLIGENCE AND DATA PROTECTION GUIDE.

https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection-2-0.pdf of the GDPR[2], not only in a new technological environment, brought about by Artificial Intelligence systems, but also in the new regulatory environment that has been built around them.

As will be observed, it is not always easy to discern which singularity in the application of the GDPR results from the approved regulatory text (AIA) and which is actually linked to the characteristics of the technology used. Therefore, in the description of any factual assumption regarding AIS -as a technology-, the assumptions that have been accepted as a matter of fact for the purposes of the AIA will be assumed, which will allow us to focus the object of study.

Having pointed out the above, our starting point is that an ethical regulation of AI involves guaranteeing, with an adequate framework of legal certainty, compliance with personal data protection standards, especially if we consider that for the fundamental right, for the principles around which its protection is structured, AIS pose specific challenges, and generate risk spaces, certain and current[3]. We were reminded in the process of drafting and approving AIA that "*data (personal and non-personal) in AI are in many cases the key premise for autonomous decisions, which will inevitably affect people's lives at various levels*".[4]

II. Artificial Intelligence systems and the processing of personal data

1. Artificial Intelligence Systems

The definition and characterisation of AIS, as an object of regulation by the GDPR, had already been considered indispensable in order to guarantee legal certainty, taking into account, however, the necessary flexibility in the continuous technological progress (Recital 12 GDPR)[5]. In our case, it will

[2] In any scenario, it should be made clear that the legal framework for the protection of personal data is not only set out in the GDPR. We will consider for these purposes Article 16 TFEU, Article 8 CDFUE; the GDPR, Regulation (EU) 2018/1725; and Directive (EU) 2016/680 (Law Enforcement Directive). In general, references to the subject matter will be made to the GDPR, avoiding adding complexity to the general reflections.

[3] Of interest, on the limitations of data protection to respond to the new needs of AI, COTINO HUESO 2022 p. 85 and following. 85 et seq.

[4] EDPS and EDPB, 2021, p. 8.

[5] It can be read in conjunction with the provisions on delegated acts of the European Commission (Recital (52), Articles 6(6), 7, 43(5) and (6), 47(5), 51, 53(6) in conjunction with Article 97 ("Exercise of the delegation") IAR.

provide us with the context of application of the GDPR, although, as it is the subject of other works in this work, we will only briefly identify those elements of the definition of CIS that we consider relevant.

Art. 3.1 AIA provides that an AIS “***is a machine-based system that is designed to operate with varying levels of autonomy and that may*** **exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives,** ***how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or*** **virtual environments”.** This definition should be completed, again in an introductory way, with the so-called general-purpose AI models, referring to *“an AI model, also one trained with with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications.*[6]*”* (Art. 3.63). Finally, the AIA refers to general-purpose AIS, as *“an AI system which is based on a general-purpose AI model and which has the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems”* (Art. 3.66).

In this definition we now highlight[7] :

- Their ability to infer understood as:

- The process of obtaining results, such as predictions, content, recommendations or decisions, which can influence physical and virtual environments.

- The ability of AI systems to derive models or algorithms from inputs or data.

- Their capacity to act with different levels of autonomy, by reference to human intervention or participation.

- Their ability to adapt after deployment or implementation and, through self-learning, to change while in use.

- It pursues explicitly or implicitly defined objectives in a specific environment, operating in a specific context.

AIS can be used “*independently or as components of a product, regardless of whether the system is physically part of the product (integrated) or contributes to the functionality of the product without being part of it (non-integrated)*”.

2. AIS lifecycle and value chain

In order to confirm and update the application of the GDPR to AIS in

[6] *“except for AI models that are used for research, development or prototyping activities prior to commercialisation”* (Art. 3.63 *in fine* AIA).

[7] Recital (12) AIA.

the context of the AIA, it is necessary to look at AIS not only from a static point of view, but also considering all the phases in their life cycle, and all their actors and agents, as well as their interactions, their *value chain*, as has been discussed by those who have addressed this issue.[8] Both concepts, life *cycle* and *value chain*, are mentioned in the AIA[9] without being precisely defined. It is clear from the AIA, and from the doctrine, that they are exponents of the complex technical reality underlying Artificial Intelligence Systems.

On the life *cycle* of AIS, the Council provided a proposal for a definition -not approved- referring to its duration from design to its withdrawal or substantial modification *(Council proposal for Art. 3.1a.)*[10]. It is worth highlighting Recital (69), which imposes the obligation to guarantee the right to privacy and the protection of personal data throughout the entire lifecycle of the AI system, even proposing technical and organisational measures with this specific task[11].

Secondly, talking about the *value chain* in AIS requires us to identify the different agents, operators, involved in its life cycle and the interactions that take place between them. This should serve as a basis for ensuring the right

[8] ("AI Watch. Artificial Intelligence for the Public Sector. Report of the "3rd Peer Learning Workshop on the use and impact of AI in public services", 24 June 2021).

[9] Recital (65); Recital (69), Recital (73); Recital (74); Recital (110); Recitals (114 and 115); Article 9.2 on risk management system; Article 12.1 on record keeping; Article 15.1 on robustness and cybersecurity accuracy; Article 40.2 on harmonised standards and standardisation documents; and ANNEX IV. Technical documentation referred to in Article 11(1).

[10] Other definitions in: AI HLEG (2020), p. 34: "*Lifecycle: The lifecycle of an AI system includes several interdependent phases ranging from design and development (including sub-phases such as requirements analysis, data collection, training, testing, integration), installation, implementation, operation, maintenance and disposal. Given the complexity of AI systems (and information in general), several models and methodologies have been defined to manage this complexity, especially during the design and development phases, such as waterfall, spiral, agile software development, rapid prototyping and incremental.*"-; AI HLEG (2019). App. 147: *"The lifecycle of an AI system encompasses the phases of development (including research, design, data provision and limited testing), deployment (including implementation) and use of that system.*"; Lazcoz and Hert, (2023), p. 8: "*Looking at the definitions in Article 3, we learn that the development phase and the use phase are the two main phases or stages of the AI lifecycle, whose key participants are the vendors and the deployers, respectively.*"; and finally, COUNCIL OF EUROPE. (2023), in its proposal for Article 10 includes in the CIS lifecycle its decommissioning (in the same sense AEPD.2020).

[11] *"In this regard, the principles of data minimisation and data protection by design and by default, as set out in Union data protection law, are applicable when processing personal data. Measures taken by providers to ensure compliance with these principles may include not only anonymisation and encryption, but also the use of technology that allows algorithms to be carried over to the data and the training of AI systems without requiring transmission between the parties or copying of the raw or structured data, without prejudice to the data governance requirements set out in this Regulation...".* See also AI HLEG, "Ethical Guidelines for Trustworthy Artificial Intelligence", April 2019.

to the protection of personal data, compliance with obligations and accountability under the GDPR[12].

From this point of view, the AIA identifies as agents or participants along the life cycle and in the value chain, among others, the *provider* (Art. 3.3 AIA), the deployer (Art. 3.4 AIA), the *authorised representative* (Art. 3.5 AIA), the *importer* (Art. 3.6 AIA), the *distributor* (Art. 3.7 AIA), *suppliers of models and AIS for general use* (Recitals (97) and (101), among many others, and Art. 53 et seq. 53 et seq. AIA), *suppliers of systems, tools, AI services, components or processes incorporated by the provider into the AIS* for, among other purposes, training, retraining, testing and evaluation of models, integration into the software or other aspects of the development of those models (Recitals (88) and (90) AIA and Art. 25 AIA). In addition, Article 25 AIA, concerning "*responsibilities along the AI value chain*" also refers to those who put their name or a trademark on the AIS, those who modify them substantially, or those who modify the intended purpose, among others.

In relation to these interactions and agents, different value chain models have been described, by way of example: the development or implementation of an internal AIS, where the provider and user coincide; customised development of an AIS for another entity; an entity writes the code and trains the system, and markets it through restricted access to the AIS, so that the user cannot make changes, only send input data and receive results; an entity sells pre-trained models and the entity that acquires the model incorporates training data; a provider sells an AIS that can be upgraded when those responsible for the deployment introduce new data; a developer of an AIS sells it to the user when those responsible for the deployment introduce new data; a developer of an AIS sells it to the user; a provider sells an updatable AIS when new data is introduced by the deployers; a developer of an AIS sells it to another AIS developer, for further training, to improve it, or to adapt it to more specific tasks – thus different data sets work; an entity integrates different AIS (e.g. the AIS decides to which AIS the user is going to use); a provider sells an AIS to another AIS developer, for further training, to improve it, or to adapt it to more specific tasks – thus different data sets work; an entity integrates

[12] It is worth recalling at this point that Recital 79 GDPR can be related to Recital (83) AIA: "*Taking into account the nature and complexity of the value chain of AI systems and in line with the principles of the New Legislative Framework, it is essential to ensure legal certainty and facilitate compliance with this Regulation. It is therefore necessary to clarify the specific role and obligations of relevant operators along the value chain, such as importers and distributors who may contribute to the development of AI systems. In certain situations, these operators could play more than one role at the same time and therefore have to cumulatively fulfil all relevant obligations associated with these roles. For example, an operator could act as both a distributor and an importer at the same time.*

different AIS (e.g. the AIS decides to which AIS it is going to use).e.g. the AIS decides to which AIS the input data is derived)[13].

This analysis of the AIS, in a value chain context, had already been considered by the AEPD (Spanish Data Protection Agency), in the identification of the processing of personal data, by reference to AI components that are included or used, and integrated in turn by other components referring to data collection, file systems, security modules, user interfaces, among others[14].

3. Processing of personal data and AIS

As a premise we must consider that "*the development and use of AI systems will in many cases involve the processing of* personal *data*"[15], which is precisely the approach of the AIA.

The processing of personal data in the life cycle of an AIS may occur at different times and in different functionalities. Indeed, as a factual scenario to be considered, different operations with defined immediate purposes are foreseen in the AIS, without prejudice to the possibility of being integrated, together with other operations, in a more complex data processing. These operations may take place during the design, development or use of AI systems[16]. Thus, we can refer to processing operations performed on so-called *training data* (Art. 3.29 AIA)[17]*, validation data* (Art. 3.30 AIA)[18]*, testing data* (Art. 3.32 AIA)[19], and *input data* (Art. 3.33 AIA)[20] provided to or acquired by the AIS, from which the AIS produces the output information. Even the existence of personal data in the algorithm itself has been described, as a consequence of the technical possibility to retrieve the data used in the training, and to make progress in the identification of the data subjects[21] .

We can also refer to the processing operation for which the AIS serves as an instrument, on the understanding that the purpose of the AIS, refer-

[13] Engler, A. C., & Renda, A. (2022).

[14] AEPD, 2020, p. 12.

[15] EDPS and EDPB, 2021, p. 14 -para. 15-.

[16] Recital (10) AIA, second paragraph.

[17] "*the data used to train an AI system through fitting its trainable parameters".*

[18] *"data used forr providing an evaluation of the trained AI system and for tuning its non-learnable parameters and its learning process in order, inter alia, to prevent underfitting or overfitting ".*

[19] *"data used to provide an independent evaluation of the AI system in order to confirm the expected performance of the AI system prior to its introduction into the market or its putting into service".*

[20] *"data provided to or directly acquired by an AI system on the basis of which the system produces an output ".*

[21] On this risk, Veale, M., Binns, R., & Edwards, L. (2018) are recommended reading. Also Whereas (76).

ring to the output information (Art. 3.1 AIA) such as contents, predictions, recommendations and even decisions, may coincide with the purpose of the processing, either as an operation or as a set of personal data processing operations.

The AIA makes numerous references to personal data, including processing, which is defined by the explicit or implicit purpose pursued. This purpose does not necessarily have to align with the lawfulness of the data under the GDPR. We speak of *biometric data* (Art. 3.34) AIA, *biometric identification and verification* (Art. 3, paragraph 35 and 36 AIA); *biometric-based data*; *special categories of personal data*; *ultra-falsification*; *profiling data*; *social score data*; *social behavioural data*; *emotions...*[22]. Similarly, we can also consider AIS as a processing operation that is integrated in a set of processing operations with a higher purpose, even if it is integrated in a processing operation with other phases or operations that do not use technological tools such as AIS.

In all these cases, the point of connection with the legal framework of protection is the existence of personal data and their processing, even if integrated in a set of non-personal data[23]. Conversely, while it is true that anonymous or anonymised data should not be considered personal data for the purposes of the GDPR[24], it should be recalled that the anonymisation of personal data is a processing operation that is subjected to the GDPR and carries its own risks. Additionally, technological developments also have an impact on the possibilities of re-identification, which must be considered [25].

We do not have space to analyse the personal data mentioned in the AIA, in any of its functionalities in the lifecycle or value chain of the AIS, or even the processing that occurs in the development or operational phase -as the

[22] Recital (18) AIA.

[23] Considering 10 AIA.

[24] Paragraph 55 et seq. of the Judgment of the CJEU (Grand Chamber) of 5 December 2023, Question referred for a preliminary ruling. Deutsche Wohnen SE and Staatsanwaltschaft Berlin Case C-683/21 (JUR 2023-432912).

[25] COUNCIL OF EUROPE (2010). "*99. The text wishes to respond to the objection raised that the recommendation goes beyond the scope of Convention 108 insofar as it covers or could cover, at least in steps 1 and 2, the processing of non-personal data, namely anonymous data. As explained in the introduction, in relation to this objection, this recommendation was intended to cover, even if only incidentally, the collection and processing of anonymous data to the extent that the processing of such anonymous data in the first and second steps may be crucial for determining the legitimacy and security of the processing in the third step, and that the three steps actually constitute a continuous process. Thus, for example, it seems unnecessary to require controllers to use accurate, authentic and up-to-date anonymous data during the first phase of data storage, especially since, prima facie and in principle, Convention 108 does not cover anonymous data. In fact, the real substance of such anonymous data may, to some extent, as a result of profiling, subsequently and unexpectedly find its way into the profile of an identified or identifiable person'.*

very purpose of the AIS-[26]. It should be remembered that one of the characteristics of AIS, as recognised by the AIA, is their capacity to *infer*[27], so that considering personal data[28], or even anonymous data -for the training of the algorithm at this point- and in view of specific input data -provided or obtained by the system- new personal data are inferred as a result.

As observed, the GDPR is particularly cautious when it comes to the processing of personal data, specifically *profiling*[29]. And this is so insofar as the result of profiling -in the sense described- whatever the technology used[30], including AI, is still personal data in itself, in respect of which full compliance with the GDPR, in all its principles and rules, must be ensured[31]. In any case, the AIA considers that profiling in the context of an AIS, beyond the risks inherent to it[32], may pose a significant risk to the health, security or fundamental rights of natural persons, which is a determining factor for its prohibition (Art. 5.1.c) and d) AIA), of its consideration as a high risk AIS (Annex III AIA[33]) and even of the caution referred to in Article 6.3, last paragraph, i.e.

[26] The debate on the application of the prohibition of processing operations contained in Article 9.1 RGPD, relating to biometric data, when "*intended to uniquely identify a natural person*", is well known. Indeed, the scope of the prohibition in Article 9(1) GDPR was disputed, depending on whether it concerned the verification-authentication of a person's identity ("one-to-one") – Art. 3(36) GDPR – or identification ("one-to-many") – Art. 3(35) GDPR. In general, the supervisory authorities consider that in both cases we are dealing with processing operations subject to the prohibition, and its exceptions, of Article 9 RGPD -Directives 5/2022 EDPS, paragraph 12 and AEPD (Nov 2023) paragraph IV.A. -. Noting the above, Recital (17) CPR in relation to verification-authentication attributes to them a likely lower impact on the fundamental rights of natural persons than remote biometric identification systems that can be used for the processing of biometric data of a large number of persons without their active participation, for the purposes of their exclusion from prohibited CIS and AR CIS, Article 5(1)(h) and 2 CPR. A revision of the above-mentioned criteria should not be linked to this statement and legal regime.

[27] For example, Recitals (12), (30), (31), among many others.

[28] High availability in terms of volume, variety and speed is important ICO UK (2019) p. 6.

[29] A fuller analysis of their specific risks (2010 Recommendation) Paragraphs 49.2 et seq.

[30] See the rules contained in the DSA to this effect, particularly Article 28.2 on the protection of minors online, preventing the presentation to minors of interface advertisements based on profiling. Also of interest is its Recital (94).

[31] Recital (10). *It should also be clarified that data subjects continue to enjoy all the rights and guarantees conferred on them by Union law, including rights related to exclusively automated individual decision-making, including profiling.*

[32] *"Whereas the lack of transparency, or even "invisibility", of profiling and the lack of precision that may result from the automatic application of pre-established rules of inference may pose significant risks to the rights and freedoms of the individual."* (COUNCIL OF EUROPE (2010) p. 5).

[33] Recital (53). *"In any case, AI systems referred to in Annex III should be considered to pose signif-*

that "*(n)otwithstanding the provisions of the first paragraph, the AI systems referred to in Annex III shall always be considered as high risk where the AI system performs profiling of natural persons...*". This is the case even if they do not support or are not a condition for the decision based solely on automated processing within the meaning of Art. 22 GDPR, in any of the interpretations that could be given to the guarantee of human intervention, as the impact on the individual is unquestionable[34]. On automated decisions under Art. 22 GDPR in relation to AIA, comments will be made in Section III.6 of this paper.

4. Identification of controllers of personal data in the AIS

One of the structural elements surrounding the protection of personal data, as a fundamental right, is the proper identification of all those involved in the processing of personal data, for the purposes of compliance with the legal requirements contained in the GDPR, with the scope and extent that, indicatively, will be set out in section III.2.

The identification of the data controller ("*alone or jointly with others*") is central to this point[35]. To this end, we can initially identify two relevant issues, which are regulated in the GDPR, and which, for the resolution of this question, have been extensively analysed by Community case law.

Thus, on the one hand, we can refer to the indication contained in Article 4.2 GDPR in the definition of ***processing*,** which includes *"operations or sets of operations which are performed on personal data or sets of personal data*". We are interested here in actions carried out on personal data that could be considered complex, consisting of different phases or stages, all of which constitute a processing of personal data[36]. This context should be completed with the definition of ***controller*** contained in Article 4.7 GDPR referring to whoever determines the purposes and means of the processing, as well as, with the

icant risks of harm to the health, safety or fundamental rights of natural persons if the AI system involves profiling within the meaning of Article 4(4) of Regulation (EU) 2016/679 and Article 3(4) of Directive (EU) 2016/680 and Article 3(5) of Regulation 2018/1725."

[34] Even if the conditions set out in Recital (53) are met, for the AIS should not be understood as having a substantial influence on the outcome of decision-making.

[35] "*The principle enshrined in Article 5(2) of the GDPR is, in our view, the most relevant principle of the GDPR, as it is intimately linked to the other six principles of the GDPR (...) Accountability requires controllers to take responsibility for what they do with personal data, to comply with all the other principles of the GDPR and to demonstrate this compliance.*" (Lazcoz and Hert, 2023, p. 20).

[36] Paragraph 72 of the CJEU (Second Chamber) Judgment of 29 July 2019, Fashion ID in Case C-40/17 (ECJ 2019, p. 148) can be analysed on this point: *"It follows from this definition -with the same content in the Directive- that a processing of personal data may consist of one or more operations, each of which relates to one of the various stages that a processing of personal data may contain*".

same introductory scope, the ***processor***, who processes *personal data on behalf of the controller (Art. 4.8 GDPR).*

From this point onwards, in our task of analysing the processing of personal data that takes place in the context of AIS, the reality that we must consider is determined on the one hand by its *life cycle* and *value chain* (Section II.2) and on the other, by the need, which is also a difficulty, to establish a framework of legal certainty. This has been highlighted by several authors[37] and during the regulatory procedure, in particular by the EDPS and EDPC in their joint report[38].

In accordance with the above, the GDPR establishes rules regarding the interactions between controllers, co-responsible parties and processors, which may serve the purpose at hand, but which may prove to be insufficient:

- Article 26 GDPR, concerning the division of responsibilities between joint controllers, imposes on them the obligation to establish in a transparent manner, by agreement -the essence of which shall be made available to data subjects- their respective responsibilities, roles and relationships, in compliance with the GDPR, in particular, with exceptions, as regards the exercise of the data subject's rights and their respective obligations to provide information, and may designate a point of contact for data subjects.

- In any event, the data subject may exercise his or her rights under the GDPR, *"with respect to and against each of the controllers*". It seems, in short, that whatever the distribution of the functions assigned to the joint controllers by agreement, this does not bind the data subject, who may exercise his or her rights (ex GDPR) against any of them.[39]

- Article 17(2) GDPR, in relation to the "*right of erasure ("right to be forgotten")* refers to the same in a responsible processing environment in terms of secondary use of personal data[40] requiring the adoption of *reasonable measures*, to inform other controllers[41]".

[37] "*For governance and accountability mechanisms to hold those responsible for the development, deployment and use of AI technologies to account for their performance and impact, the dynamics of supply chains must be urgently addressed.*" (Cobbe et al., 2023, p. 1197).

[38] *"The EDPS and the EDPS welcome the involvement of all stakeholders in the AI value chain in the regulation and the introduction of specific requirements for solution providers, as they play an important role in the products using their systems. However, the responsibilities of the different parties (user, provider, importer or distributor of an AI system) should be clearly circumscribed and allocated. In particular, when processing personal data, special attention should be paid to the consistency of these roles and responsibilities with the notions of data controller and data processor under the data protection framework, as the two rules are not congruent*". (EDPS and EDPB, 2021, p. 10).

[39] Mahieu et al., 2018, p. 52.

[40] Brown, 2023, p. 36.

[41] *"Where the controller has made the personal data public and is obliged pursuant to paragraph 1 to*

- Therefore, in an *online* environment, the GDPR imposes on the controller, in case of exercise of the right, the obligation to inform and instruct others to delete any "*links to them, or copies or replicas of such data*", and to take reasonable steps, taking into account the technology and means at its disposal, including technical measures (Recital 66 GDPR). In development and application of Article 17(2) GDPR, Article 70(d) GDPR assigns to the EDPS the task of ensuring the consistent application of this provision.

- Article 82 (*"Right to compensation and liability"*), paragraphs 4 and 5, GDPR, refers to the alleged co-responsibility, foreseeing, in case of being liable for damages caused by the processing, that each of the co-responsible parties shall be held liable for all of them in order to "*ensure effective compensation of the data subject*", without prejudice to the right of recourse over the rest, where appropriate in accordance with Art. 82.2 RGPD[42]. No similar rule is contained in Article 83 GDPR ("*General conditions for imposing administrative fines").*

Against this, difficulties have been described in the context of AIS:

- In many cases, no actor will have sufficient knowledge or control over production and deployment to be able to reliably assess or mitigate impacts and risks[43]; or even, if we consider that the right of access includes the right to know the identity of the recipient of personal data[44] allowing to know the data flow, for the purposes foreseen in the GDPR, this right would be limited to categories of recipients depending on the context of the value chain.

- Accountability in complex value chains can be made difficult as a result of their cross-border nature, despite the provisions of the GDPR[45] .

erase the personal data, the controller, taking account of available technology and the cost of implementation, shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data".

[42] This quote from Art. 82 GDPR is not intended to enter into the debate on civil liability in the context of Artificial Intelligence, and its relationship with the civil liability framework for joint controllers of personal data. On civil liability and AI of interest the document "REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL AND THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE on the security and civil liability implications of Artificial Intelligence, the internet of things and robotics", of 19 February 2020 (COM (2020) 64 final), and the proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the adaptation of the rules on non-contractual civil liability in the field of Artificial Intelligence (AI Liability Directive (COM/2022/496 final).

[43] Cobbe et al., 2023, p. 1195.

[44] Judgment of the CJEU (First Chamber) of 12 January 2023. Österreichische Post. In Case C-154/21, (ECJ 2023:4) (ECLI:EU:C:2023:3).

[45] *"While some laws, such as the EU Data Protection Act and the AIA and the California Consumer Privacy Act have sought extraterritorial effect to address regulatory arbitrage, the cross-border nature of supply*

- Lack of standardisation or common specifications, interaction between different components, hidden data interactions and imbalance in the relationship between the different actors in the value chain.

- Uncertainties in the allocation of responsibilities according to the stages in the processing operations, and varying degrees of responsibility, in the case of joint controllers, where a framework for collaboration and coordination between them has not been established[46] .

- The possible lack of documentation of actors and interactions in the lifecycle and value chain can lead to difficulties in verifying compliance with the GDPR across different personal data processing operations, even if they are connected, or form part of a set[47] .

- Obtaining consent from data subjects under the GDPR is made difficult due to a lack of interaction or of knowledge of the data subjects[48].

These being the difficulties, among others, compliance with the GDPR cannot be waived and "a broad view of supply chains, seeking to identify all operators and actors, what their role and intervention is, and how to allocate responsibilities between them[49]". Thus, those responsible for the deployment of AIS must excercise special diligence, so that they adopt the necessary measures, if applicable contractual, to ensure compliance with the GDPR in the context of the life cycle and the value chain, thereby increasing the risk posed by the integration of non-compliant *services*, not subject to accountability, or lacking the necessary documentation[50], which may even lead to the non-integration of the service.

This solution is approached by the obligation to declare compliance with the GDPR, as a content of the conformity assessment (ANNEX V ("EU Declaration of Compliance"), paragraph 5[51], AIA) when the AIS to which it refers involves the processing of personal data.

chains and enforcement difficulties remain a significant accountability challenge." (Cobbe et al., 2023, p. 1196).

[46] Mahieu et al., 2018, pp. 50 and 51.

[47] Brown, 2023, p. 19.

[48] Brown, 2023, p. 19.

[49] Cobbe et al., 2023, p. 1197. In the same vein EDPS and EDPB, 2021, p. 10: *"the responsibilities of the different parties (user, provider, importer or distributor of an AI system) should be clearly circumscribed and allocated. In particular, when processing personal data, special attention should be paid to the consistency of these roles and responsibilities with the notions of data controller and data processor as implemented by the data protection framework, as the two rules are not congruent'.*

[50] In this sense Mahieu et al., 2018, pp. 50 and 51.

[51] *"The EU declaration of conformity referred to in Article 47 shall contain all of the following information:... 5. Where an AI system involves the processing of personal data, the statement that the AI system complies with* Regulation (EU) *2016/679,* Regulation (EU) *2018/1725 and* Directive (EU) *2016/680*."

It should be recalled at this point that the case law of the CJEU and the supervisory authorities have completed the framework for the identification of controllers in complex environments, according to the following general guidelines:

- A broad interpretation of the identification of the controller must be taken as a starting point, insofar as EU law requires effective and complete protection of data subjects under the terms of the GDPR[52].

- The determination of the purposes and means of processing personal data, as a whole and each of the individual processing operations, must be analysed on a case-by-case basis, from a factual point of view. It is based on the influence on the processing of personal data and thus participating in the determination of its purposes and means can be considered a data controller[53] .

- When distinguishing between several processing phases or several sets of treatment operations, this does not refer to the different activities of the material execution of the treatment but to the existence of processing steps with different design -what data, for what purposes and by what means-[54].

- Co-responsibility does not necessarily imply that, with regard to the same processing of personal data, the various actors are equally responsible, but that they may be involved at different stages of the processing and to different degrees. The actors in this case can only be jointly and severally responsible for the processing operations whose purposes and means they jointly determine. They cannot be held responsible for operations upstream or downstream in the processing chain for which they do not determine the purposes and means[55]. Thus, the control exercised by a given entity may extend to the entire processing in question, or to a particular stage of processing.

[52] Paragraphs 34 and 66 CJEU (Grand Chamber) Judgment of 13 May 2014, Google Spain S.L. and AEPD in Case C-131/12, (ECJ 2014-85), albeit referring to Article 2.(d) of the Directive; paras 26, 27 and 28 CJEU (Grand Chamber) Judgment of 5 June 2018, Wirtschaftsakademie Schleswig-Holstein GmbH, Case C-210/16, (ECJ 2018\120) (ECLI:EU:C:2018:388); paras 65 and 66 Fashion ID Judgment (ECJ 2019\148).

[53] Paragraph 68 of the CJEU Judgment Google Spain S.L. and AEPD (ECJ 2014-85) cited; and paragraph 68 of the CJEU (Grand Chamber) Judgment of 10 July 2018, Jehovan todistajat, in Case C-25/17. (ECJ 202163) (ECLI:EU:C:2018:551).

[54] Judgment TJUE, Jehovan todistajat (TJCE 2021\63) cited, paragraph 71. Also of interest Judgment of the Spanish Constitutional Court 42/2022, of 21 March 2022. Appeal for constitutional protection 4011-2020 (RTC 2022\42).

[55] Paragraphs 70 and 74 CJEU Judgment Fashion ID (ECJ 2019\148) cited, with acceptance of the Advocate General's conclusion 101; paragraph 66 CJEU Judgment, Jehovan todistajat (ECJ 2021\63) cited . In the same sense, the recent Judgment of the CJEU (Fourth Chamber) of 7 March 2024, IAB Europe and Gegevensbeschermingsautoriteit in case C-604/22 (JUR 202473167), paragraph 78.2 (operative part).

- Co-responsibility does not imply that each of the co-responsible parties has access to the personal data concerned, if it outsources the processing activity by having a determining influence on the purpose and the essential means[56].

III. The Relationship of the Artificial Intelligence Act and the General Data Protection Regulation

1. The need for a framework for the relationship between the two bodies of law

If we focus, as we have anticipated, on the relationship between the two regulatory frameworks, we must highlight the different object and objective of the two regulations.

Indeed, the AIA concerns the placing on the market, putting into service and use of AIS in the EU, establishing prohibitions on certain AI practices, specific requirements and obligations in the case of high risk AIS, transparency rules for certain AIS, harmonised rules for the placing on the market of certain general-purpose AI models and rules on market surveillance, governance and enforcement (Art. 1 AIA). All within their scope of application (Art. 2 AIA).

For its part, the GDPR concerns the protection of natural persons, their rights and freedoms with regard to the processing of personal data, and their free movement (Art. 1 GDPR, without prejudice to its specific material scope (Art. 2 GDPR) throughout its territorial scope of application (Art. 3 GDPR).

We must remember in any case that the GDPR, although referring to the processing of personal data, to its protection as a fundamental right, really applies to the service of the individual, his freedom, his dignity, and his fundamental rights, particularly in those cases in which, as a consequence of profiling and automated decisions, his development, his possibility of self-determination or choice is conditioned, in many cases, on the basis of an inference, of a probability[57].

[56] Paragraph 38 of the ECJ Judgment Wirtschaftsakademie Schleswig-Holstein GmbH (ECJ 2018120) cited; paragraph 69 of the ECJ Judgment Jehovan todistajat (ECJ 202163) cited; paragraph 69 of the ECJ Judgment Fashion ID (ECJ 2019148) cited; and paragraph 78 (operative part) of the ECJ Judgment IAB Europe and Gegevensbeschermingsautoriteit (JUR 202473167) cited.

[57] GPA. 2020. p.3.

Therefore, in short, from this point of view, the GDPR is fully applicable to the entire life cycle of the AIS and to its entire value chain, if in its context and with whatever scope personal data are processed or, based on them, the data subject is affected by automated decisions[58], with regulatory compliance, also with regard to the GDPR, being one of the foundations of an ethical AI. Therefore, the relationship between AIS, their constitutive elements, and the processing of personal data has been considered by the doctrine[59] and the supervisory authorities in the reports issued in the course of the process of drafting the AIA.

In any case, the voices that, considering the importance of personal data in AIS and in order to avoid endangerment, directly or indirectly, the fundamental right to their protection, asked in the processing of the AIA:

- A clearly defined relationship between the proposal of the AI act and the existing data protection legislation, avoiding any inconsistency and potential conflict, avoiding any impact on or interference with it, including on the competence of supervisory authorities and governance.
- Respect the legacy of knowledge generated in the interpretation and application of Directive 95/46/EC and the GDPR, by the Article 29 Working Party, the European Data Protection Committee, the European Data Protection Supervisor, the different supervisory authorities, and the courts[60].

The legal framework for the protection of personal data is incorporated in the AIS to the extent to which it serves or supports the processing of personal data. Article 8 of the Charter of Fundamental Rights of the Euro-

[58] EDPS 2022, p.8. *"are in many cases the key premise for autonomous decisions that will inevitably affect people's lives at various levels"*. In the same sense EDPS and EDPB, 2021, p. 8.

[59] Martínez Martínez, Ricard. "ARTIFICIAL INTELLIGENCE BY DESIGN. CHALLENGES AND STRATEGIES FOR REGULATORY COMPLIANCE". *Revista Catalana de Dret Públic*, n.º 58 (2019) pp. 73 et seq.

[60] EDPS. 2022. "Opinion 20/2022 on the Recommendation for a Council Decision authorising the opening of negotiations on behalf of the European Union for a Council of Europe Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law; EDPS, and EDPB. 2021. p. 14 and p. 24 – paragraph 3, subparagraphs 56 and 57: "15. ...*It is of utmost importance to ensure clarity of the relationship of this proposal with existing EU data protection legislation. The proposal is without prejudice and complements the GDPR, the EU-PRDPA and the LED. While the recitals of the proposal clarify that the use of AI systems should continue to comply with data protection legislation, the EDPS and EDPS strongly recommend clarifying in Article 1 of the proposal that Union legislation for the protection of personal data, in particular the GDPR, the EU-PRDPR, the ePrivacy Directive 10 and the LED, will apply to any processing of personal data falling within the scope of the proposal. A corresponding recital should also clarify that the proposal is not intended to affect the application of existing EU legislation regulating the processing of personal data, including the roles and powers of the independent supervisory authorities competent to monitor compliance with those instruments."*

pean Union specifies that personal data must be processed fairly for specified purposes and on the basis of the consent of the data subject or on another legitimate basis laid down by law, that everyone has the right of access to data which has been collected concerning him or her and the right to obtain its rectification, and that compliance with these rules is subject to control by an independent authority. These requirements apply, in particular, to various provisions of the GDPR.

The legal framework for the protection of personal data, as a fundamental right, is imposed on the processing of personal data and has as its main axes:

- The establishment of requirements for the processing of personal data, which is embodied in the so-called processing principles (Article 5 and following GDPR), both for the training of AI models and for input data for the purpose of generating a prediction or inference, as examples. Of particular importance here are the principles of transparency, purpose limitation, data minimisation and accuracy, limitation of the retention period or integrity, among others.

- The recognition of data subjects' rights (Articles 12 and following GDPR), particularly in relation to automated decisions under Article 22 GDPR, without detracting from the rights of information and access (Articles 12 to 15 GDPR), the rights of rectification and erasure (Articles 16 and 17 GDPR), the right to restriction of processing (Article 18 GDPR) and the right to object to processing (Article 21 GDPR), all of which relate to personal data processed at any point in the life cycle of the AIS.

- The establishment of obligations for those involved in one way or another in the processing of personal data. These obligations include data protection by design and by default (Art. 25 GDPR), and the Data Protection Impact Assessment (Art. 35 GDPR).

Of particular importance, it is also imposed on all those responsible, not only regarding to processing in high risk AIS, the obligation to implement, review and update appropriate technical and organisational measures to ensure and be able to demonstrate that the processing is in compliance with the GDPR, taking into account the nature, scope, context and purposes of the processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons (Art. 24.1 GDPR).

Similarly, the notification and management of security breaches of personal data referred to in Articles 33 and 34 of the GDPR is also an obligation of all data controllers:

- As a result of Article 5.2 (GDPR) the controller must be able to demonstrate compliance with the obligations assumed under the GDPR ("*proactive accountability*").

- The establishment of an institutional framework of guarantees, highlighting in this respect the intervention of *independent data protection supervisory authorities,* in an environment of legal certainty and security.

A key element in this framework of guaranteeing the fundamental right is the identification of the controller of personal data, even in cases where the controller determines the purposes and means of the processing jointly with other controllers, or where the processing is carried out on behalf of a controller, for example, by a processor acting on its behalf. It has thus been considered indispensable for the protection of the rights and freedoms of data subjects, for the clear attribution of responsibilities, also for the supervision by supervisory authorities[61].

2. Application of the Regulation without prejudice to the application of the GDPR

A basic assumption considered by the AIA is the full application of the GDPR to the AIS, exclusively and to the extent that it involves or determines a processing of personal data, or an automated decision, falling within its scope. From this point of view, Recital 15 GDPR is consistent with and can be analysed according to which *"In order to prevent creating a serious risk of circumvention, the protection of natural persons should be technologically neutral and should not depend on the techniques used ".*

This general statement has also been provided for in the AIA. Thus, Recital (10) states: *"...This Regulation does not seek to affect the application of existing Union law governing the processing of personal data, including the tasks and powers of the independent supervisory authorities competent to monitor compliance with those instruments...*". The same Recital reminds us that the AIA "*does not affect the obligations of providers and deployers of AI systems in their role as data controllers or processors stemming from Union or national law on the protection of personal data in so far as the design, the development or the use of AI systems involves the processing of personal data.", and goes on to clarify that "data subjects continue to enjoy all the rights and guarantees awarded to them by such Union law, including the rights related to solely automated individual decision-making, including profiling", and concludes that 'harmonised rules for the placing on the market, the putting into service and the use of AI systems established under this Regulation should facilitate the effective implementation and enable the exercise of the data subjects' rights and other remedies guaranteed under Union law on the protection of personal data and of other fundamental rights".*

From these statements, without prejudice to various Recitals with the same content[62]*, Article 2(7) AIA states: "Union law on the protection of personal data, privacy and the confidentiality of*

[61] See in this respect Recitals 13 and 79 GDPR.

[62] Recitals (95), (157) – on independent personal data protection supervisory authorities.

communications applies to personal data processed in connection with the rights and obligations laid down in this Regulation. This Regulation shall not affect Regulation (EU) 2016/679 or (EU) 2018/1725, or Directive 2002/58/EC or (EU) 2016/680, without prejudice to Article 10(5) and Article 59 of this Regulation". The meaning and scope of the last subparagraph will be discussed in the following sections.

By way of example, more precisely, in relation to *remote biometric identification* for the *targeted search of a convicted or suspected offender*, the framework of obligations contained in Art. 29(10) for those responsible for the deployment of high risk AIS is "*without prejudice to Article 9 of Regulation (EU) 2016/679 and Article 10 of Directive (EU) 2016/680 for the processing of biometric data".*

Finally, we must consider the calls that the AIA makes to the precepts of the GDPR (of the Regulation (EU) 2016/679 (LED) and Regulation (EU) 2016/679) to define or complete the definition for its own purposes, being able to highlight the definition of personal and non-personal data, special categories of personal data or profiling[63].

3. The Regulation as "lex specialis" for policing purposes

The AIA contains among its provisions rules that complete the regulatory framework of the right to the protection of personal data as a fundamental right, and this is precisely as a consequence of the specific challenges that AI poses for them. In accordance with the foregoing, the AIA, already from the Commission's proposal, is based in some of its provisions on Article 16 TFEU and Article 8.1. of the EU Charter of Fundamental Rights, proposing from this point of view complementary rules to the GDPR.

In this sense, the European Commission's initial proposal can be read as being without prejudice to the GDPR and the LED, *"which it complements with a set of harmonised rules applicable to the design, development and use of certain high-risk AI systems and with restrictions on certain uses of* remote *biometric identification systems*"[64].

It refers in particular to the prohibition, subject to exceptions, of the use of AIS for real-time remote biometric identification in publicly accessible

[63] Art. 3(37) and 50 to 53 AIA.

[64] EUROPEAN COMMISSION 2021. p.4. This is also stated in Recital (3) AIA: *"...Insofar as this Regulation contains specific rules on the protection of natural persons with regard to the processing of personal data relating to restrictions on the use of AI systems for remote biometric identification for law enforcement purposes, for the use of AI systems for risk assessment of natural persons for law enforcement purposes and for the use of AI systems for biometric categorisation for law enforcement purposes, it is appropriate to base this Regulation, distant as it relates to those specific rules, on Article 16 TFEU. In the light of those specific rules and the use of Article 16 TFEU, the European Data Protection Board should be consulted.*

areas for law enforcement purposes, for the use of AI systems for risk assessment of natural persons for law enforcement purposes and for the use of AI systems for biometric categorisation for law enforcement purposes. It is regulated in this manner, which exhaustively outlines the use and data processing involved, as supported by article 16 TFEU. It must be considered as *lex specialis* in compliance with the rules on data processing contained in article 10 LED.

The scope of this *lex specialis* is the specifically defined one, not extending to other similar processing for purposes other than the application of the Law, even by competent authorities[65]. In short, the establishment of prohibitions, or even additional requirements resulting from the classification of AIS, in view of the risk it poses to the protection of personal data, should be seen as complementary to the existing data protection legal framework, in addition to the requirements and obligations set out therein (LED).

These points should be borne in mind:

- In no case can it be considered as providing a legitimate basis for the processing of personal data under Article 8 LED[66]. In general, the rules delimiting the existence of a prohibited AIS or High-Risk AIS cannot be confused with the legitimate basis for processing[67].

- This legitimate basis for processing seems to be referred to in Art. 5(5) AIA, with reference to state rules, which are to be understood as bound, not only by Art. 5 AIA, but also by the conditions imposed by the legal framework for the protection of personal data.

- The application of the GDPR must be ensured, to the extent of the risk they pose to the protection of personal data, and the principles of necessity, proportionality, purpose limitation, among others, must be observed.

- The obligation to ensure the existence of an independent authority to monitor compliance with these rules is updated[68].

[65] Recital (39): *"...In the implementation of Article 9(1) of Regulation (EU) 2016/679, the use of remote biometric identification for purposes other than law enforcement has already been subject to prohibition decisions by national data protection authorities".*

[66] Recital (38) AIA.

[67] Recital (63) AIA, and the aforementioned Recital (38). Also CEPS, 2022, p. 14*: 'Recital 41 of the proposal states that operators of AI systems must comply with the EU data protection regime, stating in particular that the risk-based categories of the AIA should not be interpreted as providing legal grounds for processing personal data. This provision provides the basis for compatibility between the AI Act and the GDPR, but its generality requires further specification of the rules in both acts, as the following sections demonstrate."* -recital 41 of the proposal is recital 63 of the adopted text-.

[68] EDPS and EDPB. 2021 p. 22 – paragraph 49. Also relevant here is the report in paragraph 50: *"However, there is no explicit provision in the proposal that assigns competences to ensure compliance with these rules to the control of independent authorities. The only reference to data protection supervisory*

4. Specific cases of legal basis for the processing of personal data in the Regulation

Thirdly, we must consider the existence of rules in the AIA that expressly introduce a legitimate basis for processing in accordance with the provisions of the GDPR, as anticipated by Article 2(7) AIA:

Article 10.5 Regulation, "to ensure bias detection and correction in relation to the high-risk AI systems".

Article 10 AIA, on data and data governance, integrated in Section 2, Chapter III ("Requirements for High-Risk AI Systems") provides in its paragraph 5 the necessary legitimate basis for the processing of personal data in accordance with Art. 6 GDPR, and lifts the prohibition referred to in Art. 9.1 GDPR, Art. 10 LED and Art. 10.1 Regulation (EU) 2018/1275. To this end, the processing, which is exceptionally enabled, must be strictly necessary to ensure the detection and correction of the negative biases associated with the high risk AIS, with the providers of such systems being considered for these purposes as data controllers.

This being so, and in accordance with Article 9 GDPR, the obligation is imposed to adopt the necessary safeguards, expressly citing the establishment of technical limitations to the re-use and the use of the most recent security and privacy protection measures, such as pseudonymisation or encryption, if anonymisation could significantly affect the objective pursued. In this respect, the initial wording was completed (Amendment 290 Proposal for a regulation Article 10 -paragraph 5), highlighting its exceptional nature, referring to negative bias and incorporating additional safeguards[69]**.** It is required, cumulatively:

- Detection and correction of bias cannot be effectively accomplished by processing other data, including synthetic or anonymous data.

- The processing of special categories of data is subject to technical limitations

authorities competent under the GDPR, or LEDs, is in Article 63(5) of the proposal, but only as 'market surveillance' bodies and, alternatively, with other authorities. The EDPS and the EDPS consider that this creation does not ensure compliance with the requirement of independent supervision laid down in Article 16(2) TFEU and Article 8 of the Charter'.

[69] *"At the same time, Article 10(5) of the proposal states that 'providers of such systems may process special categories of personal data'. Moreover, the same provision requires additional safeguards, also giving examples. Therefore, the proposal seems to interfere with the application of the GDPR, the LED and the EUDPR. While the EDPS and the EDPB welcome the attempt to establish adequate safeguards, a more consistent regulatory approach is needed, as the current provisions do not seem sufficiently clear to create a legal basis for the processing of special categories of data, and need to be complemented by additional protection measures which still need to be assessed. Moreover, where personal data have been collected through processing within the scope of LED, possible additional safeguards and limitations resulting from national transpositions of LED should be taken into account*" ([EDPS and EDPB, 2021, p. 28]).

for re-use and to state-of-the-art security and privacy protection measures, including pseudonymisation.

- The personal data processed shall be protected, subject to appropriate safeguards, including strict controls and documentation of access, to prevent misuse and to ensure that only authorised persons have access to it with appropriate confidentiality obligations, and must not be transmitted, transferred or otherwise made accessible by third parties.

- The personal data processed are deleted once the bias has been corrected or at the end of their retention period, whichever comes first.

Finally, the processing of personal data in this case requires a specific provision in the *records of the processing activities*[70] which have to include a justification why the processing of special categories of personal data was strictly necessary to detect and correct bias and this objective could not be achieved by processing other data.

It should be borne in mind, in any case, that this precept also serves the fundamental right to data protection, and the principles on which it is based[71].

Article 59 Regulation, concerning the "further processing of personal data for developing certain AI systems in the public interest in the AI regulatory sandbox".

Art. 59.1 AIA provides the legitimate basis, and lifting of the prohibition, necessary for the further processing of personal data for the development, training and testing of certain AIS, for reasons of public interest, in the controlled AI test space, in accordance with the provisions of Article 6(4) GDPR and Article 9(2)(g) GDPR[72].

We refer to AIS developed to safeguard a substantial public interest, in the areas referred to in 59.1.a) AIA, such as public safety and public health, including disease detection, diagnosis, prevention, control and treatment and improvement of health systems; a high level of protection and improvement of the quality of the environment, protection of biodiversity, pollution, as well as ecological transition, climate change mitigation and adaptation; energy

[70] Article 30 *("Register of processing activities")* GDPR.

[71] *"In order to ensure fair and transparent processing in relation to the data subject, (...).) the controller should use appropriate mathematical or statistical procedures for profiling, implement appropriate technical and organisational measures to ensure, in particular, that factors giving rise to inaccuracies in personal data are corrected and that the risk of errors is minimised and the security of personal data is ensured in a way that takes into account potential risks to the interests and rights of the data subject and avoids, inter alia, discriminatory effects on natural persons (...)." (FRA EU, 2018, p. 7).* In the same sense COTINO HUESO 2023. p. 303.

[72] *"the processing is necessary for reasons of essential public interest, on the basis of Union or Member State law, which must be proportionate to the aim pursued, respect in substance the right to data protection and provide for appropriate and specific measures to protect the interests and fundamental rights of the data subject*".

sustainability; safety and resilience of transport and mobility systems, critical infrastructures and networks; efficiency and quality of public administration and public services.

At this point, it must be taken into account that the authorisation is based on the fulfilment of cumulative requirements, together with the substantial public interest purpose referred to in Art. 59.1.a), listed exhaustively in Article 59.1 AIA, among which, from the point of view of the protection of personal data, the most important are[73] :

- The data processed are necessary to meet one or more of the requirements for high risk AIS (Chap. III, Sect. 2), where they cannot be effectively met by processing anonymised, synthetic or other non-personal data.
- Effective monitoring mechanisms are in place to determine the need for a Privacy Impact Assessment in accordance with Art. 35 GDPR.
- Personal data are in a functionally separate, isolated and protected data processing environment under the control of the potential provider and accessible only to authorised persons.
- The data has been collected in accordance with the GDPR, and cannot be shared outside of the controlled testing environment.
- The processing of the data may not give rise to measures or decisions affecting the data subjects or their rights under the GDPR, as clarified by Recital (140)[74] AIA.
- Personal data shall be protected by appropriate technical and organisational measures and shall be deleted once the participation has ended or the personal data have reached the end of their retention period.
- Records of the processing of personal data shall be kept for the duration of participation, unless otherwise provided for by Union or national law.

In any case, and from the point of view of compliance with the GDPR, the obligations imposed therein on data controllers and the rights of data subjects will be applicable, resolving the doubt raised during the processing of the proposal as to whether we were effectively in a personal data processing framework in which the scope of the obligations of data controllers and the rights of data subjects were being limited[75].

On the other hand, this space being a concrete and controlled frame-

[73] Parliament's Amendment 506 et seq. to Article 54 increased the legal, technical and organisational safeguards for the protection of personal data. In this respect, EDPS and EDPB. 2021, points 64 et seq.

[74] "*In particular, this Regulation should not provide a legal basis within the meaning of Article 22(2)(b) of Regulation (EU) 2016/679 and Article 24(2)(b) of Regulation (EU) 2018/1725.*"

[75] Recital (140) AIA; and EDPS and EDPB. 2021 paragraph 64.

work established by a competent authority and offered by that authority to providers or potential providers of AI systems to develop, train, validate and test the AIS according to a plan, on a temporary basis and under regulatory supervision[76], questions may arise about the accountability framework under the GDPR, in particular about the identification of the controller, or joint controllers, of the processing operations[77].

Finally, it should be noted that the legal basis under consideration is not foreseen in relation to the processing of personal data that might take place in the case **of '***Testing of high-risk AI systems in real world conditions outside AI regulatory sandboxes,* as provided for in Article 60. At this point, it should be considered that the requirements for processing operations, the obligations of data controllers, the rights of data subjects provided for in the GDPR fully overlap with the cumulative conditions set out in paragraph 4, in particular paragraphs e) g) h) i) and k), and paragraph 5. This consideration allows us to rule out the *informed consent* of the *data subject* (Article 61 AIA) as the consent referred to in Articles 6 and 9.2 RGPD, legitimate basis for processing or assumption of lifting of the prohibition of processing, being thus confirmed by Recital (141) *in fine AIA*[78] .

5. Independent personal data protection supervisory authorities

As indicated in Section III.1, the legal framework for the protection of personal data, as a fundamental right, includes its institutions of guarantee, highlighting in this sense the intervention of the *independent data protection supervisory authorities,* for which a framework of certainty and legal certainty is required, in the terms indicated. This intervention in the case of AIS, on the other hand, must take place within a framework of *"structured and institutionalised cooperation*" with other competent authorities[79], such as market surveillance authorities, always respecting that independence[80].

This being the starting point, we can point out:

- The AIA should be understood without prejudice to the application of the GDPR, which includes the competences, functions and powers of the independent supervisory authorities (arts 55 and following GDPR) and which

[76] Art. 3.55 AIA.

[77] This was made clear by EDPS and EDPB. 2021 paragraph 65.

[78] *"...The consent of data subjects to participate in such tests under this Regulation is separate from and without prejudice to the consent of data subjects to the processing of their personal data under the relevant data protection legislation...".*

[79] EDPS. 2022, p.14.

[80] Brown, 2023, p. 69.

also enables access to any documentation, in the understanding that it extends to the safeguard procedure to ensure adequate and timely implementation of the AIS that present a risk to fundamental rights[81].

Accordingly, Article 77 AIA on the *powers of authorities protecting fundamental rights* empowers them to request and access any documentation created or maintained under the AIA where necessary "*for effectively fulfilling their mandates within the limits of their jurisdiction*", informing the market surveillance authority. In the cases provided for in paragraph 3, they may even make a reasoned request to the market surveillance authority to organise high risk AIS testing.

These provisions must in any case be interpreted in the sense most consistent with the independence of data protection supervisory authorities and the rules governing their activities under the GDPR.

- In the context of the consideration of the AIA as *lex specialis* in relation to *"real-time" remote biometric identification systems in publicly accessible areas* for law enforcement purposes**,** without prejudice to the provision on the involvement of an independent supervisory authority, it also provides for the necessary notification to data protection supervisory authorities (Art. 5(4), (5) and (6) AIA), to be complemented by the obligations of the deployers with regard to the use of remote biometric identification systems to make available and report, in accordance with Art. 26(10) AIA[82].

The exclusion of the communication of *sensitive operational data* should be seen in conjunction with Recital (159) *in fine* according to which "*no exclusion on disclosing data to national data protection authorities under this Regulation should affect the current or future powers of those authorities beyond the scope of this Regulation*".

- Reference is made to data protection supervisory authorities in Recital (140), providing for cooperation with competent authorities in the controlled area of AI testing.

This provision is further elaborated in Article 57(10) AIA in relation to their possible involvement[83], and where they have provided guidance for compliance with the GDPR (paragraph 12).

To the extent that it prevents the imposition of administrative fines on those responsible who have followed the guidelines in *good faith* (Paragraph 1, h) *in fine*), a concept that may generate uncertainty, it will be necessary to

[81] Recitals (10) and (157).

[82] Recital (36).

[83] *"National competent authorities shall ensure that, to the extent the innovative AI systems involve the processing of personal data or otherwise fall under the supervisory remit of other national authorities or competent authorities providing or supporting access to data, the national data protection authorities and those other national or competent authorities are associated with the operation of the AI regulatory sandbox and involved in the supervision of those aspects to the extent of their respective tasks and powers.*".

establish a complete and precise regulation for the guidelines indicated, and their framework for compliance.

- Article 74(8) AIA provides for the possibility of appointing Data Protection Supervisory Authorities as market surveillance authorities in respect of the high risk AIS listed in Annex III, point 1, to the extent that the systems are used for law enforcement purposes and for the purposes listed in points 6, 7 and 8 of the same Annex.

This being a proposed scenario, it requires in any case a proper delimitation of the competences, functions and powers of these authorities in both areas, taking into account the applicable rules.

- Finally, Article 85 refers to the right to lodge a complaint with a market surveillance authority by anyone who considers that there has been an infringement of the AIA[84].

This complaint should be without prejudice to the right to lodge complaints with the data protection supervisory authorities in accordance with Article 77 GDPR, which will deal with them in accordance with Article 57.1.f) GDPR and resolve them, where appropriate, in accordance with the powers provided for in Article 58.2 GDPR -sanctioning, corrective, precautionary...-. This is the result of Article 85.1 AIA, interpreted in the light of Recital (170)[85].

It should be recalled, on the other hand, the scope of the purpose of this Article 85 claim, as according to its second paragraph, it shall be taken into account for the purpose of carrying out market surveillance activities and managed in accordance with the specific procedures established by the market surveillance authorities (Art. 11 Regulation (EU) 2019/1020)[86].

[84] At this point it should be recalled that Article 110 (Chapter XIII. Final Provisions) amends Annex I of Directive (EU) 2020/1828 of the European Parliament and of the Council of 25 November 2020 on representative actions for the protection of the collective interests of consumers and repealing Directive 2009/22/EC (OJ L 409, 4.12.2020, p. 1), including paragraph 68) which enables *representative actions* (for the protection of collective consumer interests) in case of infringement of the AIA. It should be recalled that such actions are also possible in case of infringement of the GDPR (paragraph 56) of the Directive. See COTINO HUESO (2022) pp. 83 and 84.

[85] *"Union and national law already provide effective remedies to natural and legal persons whose rights and freedoms are adversely affected by the use of AI systems. Without prejudice to those remedies, any natural or legal person that has grounds to consider that there has been an infringement of this Regulation should be entitled to lodge a complaint to the relevant market surveillance authority ."*.

[86] Reflections on collective actions may be of interest.

6. Human surveillance and automated individual decisions: Article 22 GDPR and the Regulation

Article 22 GDPR enshrines the right of every data subject "*not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her*".

Although Art. 22 GDPR can be understood as a restricted case ("... *only*..."), it is true that the latest case law interpretations analyse the relationship between the result of the automated processing and the decision taken from a qualitative point of view[87], depending on the actual influence that may be exercised[88]. This is without prejudice to the human intervention referred to in Article 22(3) GDPR of a reactive nature.

If we analyse the requirements of high risk AIS (Chapter III, Section 2, AIA) in particular the *human oversight* referred to in Article 14*,* we can conclude that the factual scenario of Article 22 GDPR, in particular in cases where the automated individual decision would be possible because it falls under one of the cases of its paragraph 2, could not take place.

The scope of Article 14 AIA is complemented by the information to be provided to the deployer according to Article 13(3) on instructions for use (d). It is also reported to ANNEX IV, paragraph 2.e), as part of the technical documentation of the high risk AIS referred to in Article 11 AIA and forms part of the content of the Fundamental Rights Impact Assessment (Art. 27.1.e) AIA, and must be incorporated into the Privacy Impact Analysis (PIAC) in accordance with the provisions of Art. 35.7.d) GDPR.

Indeed, it is questionable whether the effective compliance with the requirements imposed on high risk AIS by Article 14(4) and (5) of the abovementioned Article 14 AIA preclude the existence of individual decisions within the meaning of Article 22, which could be allowed under Article 22(2).

7. The right to explanation. Individual decisions in the context of certain high risk AIS and Article 22 GDPR

Article 86 ("*Right to explanation of individual decision-making*") establishes the right of any person affected by a decision taken by the deployer on the

[87] The Judgment of the CJEU (First Chamber) of 7 September 2023, OQ and Land Hessen, with SCHUFA Holding AG, Case C-634/21, paragraphs 53 to 55 (ECJ 2023-146), and the analysis made by COTINO HUESO (2024) can be analysed.

[88] In the same sense WG Art. 29 (2018) para. 4.1.

basis of the output of a high-risk AI system (except Annex III.2[89]), and which produces legal effects or significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken[90].

This provision is related to Article 26.11 AIA, which imposes on those responsible for deployment[91] who make decisions or assist in making decisions relating to natural persons, the obligation to inform them that they are subject to the use of the high-risk AI system[92], and this, in turn, to Article 13.1 and 3.b) iv) and v) AIA on transparency obligations and provision of information to deployers, precisely for the fulfilment, inter alia, of those obligations[93], all without prejudice to the provisions relating to the EU Database for high-risk AI system, in accordance with Article 71 AIA -in its cross-reference to ANNEX VIII, Section A, paragraph 6-[94].

It is important to establish the relationship of Article 86 AIA, and the right it establishes, with the rights of the data subject in the context of Article 22 GDPR ("*Automated individual decisions, including profiling*") and concordant on the right to an explanation in these cases, insofar as the right referred to in Article 86 "*shall only apply insofar as...it is not already provided for in Union law"* as set out in its paragraph 3.

We must consider that Article 86(1) AIA assumes that *human surveillance* of the high risk AI system has taken place -as foreseen in Article 14 AIA-

[89] *"AI systems intended for use as safety components in the management and operation of critical digital infrastructures, road traffic and the supply of water, gas, heat and electricity.*

[90] *"unless and to the extent that this obligation is restricted by Union or national law in accordance with Union law (paragraph 2)*".

[91] Without prejudice to the provisions of Article 50 on "*Transparency obligations for providers and deployers of certain AI systems"*, interacting with natural persons; or to the specific provisions relating to AR CIS for law enforcement purposes (Art. 13 CSP).

[92] For high-risk AI systems used for law enforcement purposes, Article 13 of Directive 2016/680 shall apply.

[93] On this point the European Parliament proposed the introduction of a second, more explicit paragraph: "*Transparency shall thus mean that, at the time the high-risk AI system is placed on the market, all available technical means are used in accordance with generally recognised technological state of the art to ensure that the results of the AI system are interpretable by the provider and the user. The user shall be enabled to understand and use the AI system appropriately by knowing in general how the AI system works and what data it processes, thus enabling him to explain the decisions taken by the AI system to the person concerned in accordance with Article 68(c).*"

[94] "In respect of high-risk AI systems to be registered in accordance with Article 49(1), the following information shall be provided and duly updated:...6. A simple and concise description of the information used by the system (data, inputs) and its operating logic".

which has not prevented the deploying officer from taking the decision based on the outcome of the high risk AI system.

On the interpretation of "*decision taken by the deployer on the basis of the results of an AI system"*, we could understand it to include a decision where the output information has a *significant influence*[95]. This interpretation results from Recital (171)[96] *("...where the decision...is mainly based...")*, from the empowerment of the Commission contained in Article 7.1 AIA, in conjunction with Article 7.2.g) AIA[97], and is the most favourable to the rights of those affected. It also corresponds to the evolution of the CJEU's doctrine (Note 61).

On the other hand, we have already indicated the debate on the difficulty of occurrence of the factual assumption of Article 22 RGPD, in its paragraph 2, in the case of high risk AI system, insofar as human supervision is required for these. If we were to consider that the human supervision required by Art. 14 AIA is not an obstacle to automated decisions within the meaning of Article 22 GDPR, in the cases of its paragraph 2, we could analyse:

- Article 86(1) AIA refers to decisions taken by the controller on the basis of the output of certain high-risk AI systems, in the broad terms we have indicated, and Art. 22 refers to "*decision based solely on automated processing, including profiling*", being also subject to a broad interpretation.

- The right of explanation provided for in Article 86 of the CPR refers exclusively to high-risk AI systems[98] (except for Annex III.2), and is therefore, in a way, more restrictive than the alleged fact of Article 22 of the GDPR.

- On the scope of the right of data subjects, Article 86 refers to request-

[95] See proposal of the European Parliament, for the purpose of categorisation as high risk AI system, example, Amendments 46 regarding Recital 36, Amendment 47 regarding Recital 37 and Amendment 814 regarding ANNEX III.4(a) and to some extent Recital (53) AIA.

[96] *"The persons concerned should have the right to obtain an explanation where the decision of a deploying officer is based primarily on the results of certain high-risk systems falling within the scope of this Regulation and where that decision produces legal effects or similarly significantly affects such persons in a way that they consider that it has a negative impact on their health, safety or fundamental rights. Such an explanation should be clear and meaningful and provide a basis for the persons concerned to exercise their rights. The right to obtain an explanation should not apply to the use of AI systems for which exceptions or restrictions arise under national or Union law, and should only apply to the extent that this right is not already provided for in Union law.*

[97] "*the extent to which persons who might suffer such harm or negative repercussions are dependent on the outcome generated by an AI system, in particular because, for practical or legal reasons, it is not sensible to be able to forego such an outcome;*". Also from the content of Recital (53) in setting out the conditions for identifying the substantial nature of the influence of the AIS on decision-making, in one of the human scenarios.

[98] It should be recalled that the extent to which the information in the exit information is linked to the potentially prejudicial decision is one of the elements considered for the determination of the high-risk AI systems under Article 7.1 AIA, as set out in Article 7.2(e) AIA, and was taken into account in its inclusion in ANNEX III.

ing from the deployer *"clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken"* so as to provide a basis for them to "exercise their rights" (Recital 171). It should in any case be complemented by Article 26.11 AIA, concerning the obligation to inform data subjects who are subject to the use of the high-risk AI system.

- For its part, the controller assumes the obligation to inform the data subject in a concise, transparent, intelligible and easily accessible form, in clear and plain language of any information specifically addressed to a child (Art. 12.1 GDPR). Specifically in case of automated decisions, including profiling, as referred to in Articles 22(1) and 4 GDPR (Art. 13(2)(f) and 14(2)(g) GDPR, either at the time of collection of the data, within a reasonable period of time, or before the start of the processing, it must provide information on its existence and meaningful information on the logic applied, as well as the significance and expected consequences of such processing for the data subject. As far as we are concerned, this same information must be provided, inter alia, in the case of the exercise of the right of access referred to in Article 15(1)(h) GDPR.

In view of the above, we may consider that the rights attributed to data subjects in the GDPR are broader in their assumptions, so that the exception contained in Article 86(3) of the GDPR will generally be activated in the case of automated decisions affecting individuals, especially and to the extent that it covers cases of AI systems other than high risk ones.

If the possibility of applying the case provided for in Article 22.2 GDPR to the high-risk AI system is accepted, and it is interpreted that some of the information provided for in the AIA is not provided under the GDPR, it could be requested by the data subject in the part not provided for. This would be particularly relevant for the case of application of the case referred to in Article 22(2)(b) GDPR, which is not covered by the right referred to in Article 22(3) GDPR.

It would be desirable, in any case, that the information referred to in the aforementioned art. 86 AIA should serve to construct a standard of information to be provided to the data subject in accordance with the GDPR, being incorporated into it in any case[99]. It should be remembered that, in both cases,

[99] EDPB (2022): '119. Article 15(1)(h) provides that every data subject should have the right to be informed, in a meaningful way, inter alia, about the existence and underlying rationale of automated decision-making, including profiling of the data subject, and about the significance and intended consequences of such processing69. If possible, the information pursuant to Article 15(1)(h) should be more specific in relation to the reasoning leading to specific decisions concerning the data subject who requested access." pp. 39 and 40.

69 See, in this connection, the Guidelines on Transparency under Regulation 2016/679

the information must be sufficient for the exercise of the rights to which they are entitled (Recital (171) AIA and Art. 22.3 GDPR).

8. Collaboration of the Regulation in compliance with the GDPR

The last indent of Recital (10) AIA states that *'harmonised rules for the placing on the market, the putting into service and the use of AI systems established under this Regulation should facilitate the effective implementation and enable the exercise of the data subjects' rights and other remedies guaranteed under Union law on the protection of personal data and of other fundamental rights '.*

This general statement allows us to approach the extent to which the harmonised rules contained in the AIA facilitate the application of the GDPR to AI systems when they involve, form part of, are employed in or for the processing of personal data, in the terms and to the extent that we have anticipated.

To this end, we can identify precepts of the AIA that impose certain obligations on operators and agents, mainly providers and deployers, to facilitate compliance with the GDPR. Thus:

- First, naturally, we would have the list of AI systems that are to be considered prohibited, or limited, because they involve a purpose that would also be considered prohibited under the GDPR. In this way, the AIA, and its institutional control and enforcement mechanisms, reinforce the application of the GDPR.

- The establishment of requirements to be met by high-risk AI systems (Section 2 of Chapter III) as well as the establishment of obligations for providers and those responsible for the deployment of high-risk AI systems, can also facilitate compliance with the GDPR -in the terms already indicated, including the principle of transparency- and, above all, the demonstration of compliance within the framework of *proactive responsibility* (art. 5.2 GDPR).

We are referring to compliance by providers (Art. 16), deployers (Art. 26) and other actors (Art. 22 to 25 AIA) with the obligations set out in Sections 2 and 3 of Chapter III, relating to the *risk management system* (Art. 9 AIA), *data and data governance* (Art. 10 AIA), on *technical documentation* -including the content of ANNEX IV to which reference is made and *record keeping* (Arts. 11 and 12), on *transparency and provision of information to deployers* (Art. 13 AIA), on

(WP 260), paragraph 41, with reference to the Guidelines on automated individual and collective decision-making (WP 260), paragraph 41, with reference to the Guidelines on automated individual and collective decision-making (WP 260), paragraph 41, with reference to the Guidelines on automated individual and collective decision-making (WP 260), paragraph 41.

accuracy, robustness and cyber-security (Art. 15 AIA) on the *quality management system* (Art. 17 AIA), *retention of documentation* (Art. 18 AIA), *archiving of automatically generated records* (Art. 19 AIA), *corrective measures and duty of information* (Art. 20), *responsibilities along the value chain* (Art. 25 AIA), on the *Impact Assessment on fundamental rights* (Art. 27RIA), and even compliance with the registration obligations referred to in Article 49, in conjunction with Article 16(i) AIA. The same can be said in the case of general use AI models, in accordance with Articles 51 and following AIA, as well as the content of ANNEX XI and following.

- On the data protection impact assessment, as referred to in Article 35 GDPR, the AIA contains rules that enhance its value.

Thus, according to Article 26(9) AIA, high-risk AI system controllers shall use the information provided to them in accordance with Article 13 AIA for the preparation of the data protection impact assessment (Art. 35 GDPR). Similarly, Article 27 AIA, which requires high-risk AI system deployers to prepare a fundamental rights impact assessment, states in paragraph 4 that if any of the obligations set out in the provision are already fulfilled as a result of the PIA, the fundamental rights impact assessment will complement the data protection impact assessment.

Finally, a summary of the PIA is included in relation to the information to be submitted to the high-risk AI system Register, in accordance with Art. 49(4) (ANNEX VIII, Section B)) of the GDPR, which will facilitate its knowledge, the exercise of rights under the GDPR by data subjects, as well as, where appropriate, complaints to the data protection supervisory authorities.

These statements, however, may be understood as delimiting the area of diligence required of the controller in the preparation of the PIA, and may even introduce confusion as to the obliged agent, predetermining the status of controllers in the context of the AI system. In any event, the requirement for the PIA, its content and the controller, must be determined on a case-by-case basis in accordance with the GDPR.

- Regarding the framework established around the provision of harmonised standards, common specifications, conformity assessment and certificates referred to in Section 5, Chapter III, it is clear that as they are aimed at verifying compliance with the provisions of Section 2, they can facilitate the implementation of the GDPR, even more so as they will facilitate a standardised framework for interaction between those involved, and even the identification of controllers and joint controllers of personal data in the value chain.

Although ANNEX V(5) provides, by reference to Article 47(2) AIA, for a statement drawn up and signed by the provider of the high-risk AI system, concerning the existence of personal data processing and compliance with the GDPR in that case, it is not defined as a specific task of such a conformity assessment and is not documented, allowing for confusion about the actual compliance with Union law as regards the protection of personal data in high-risk AI system, which have gone through the process, as a content of the Conformity Assessment. This provision was incorporated by the European Parliament in its Amendment 867.a. to ANNEX V and can be seen in connection with the statement contained in Recital (69), also introduced by the European Parliament.

Specifically, it would have been desirable to provide a more detailed more explanation in the operative part the statement contained in the aforementioned Recital (69)[100], and Article 10 AIA regarding data and their governance in this context, and even to give it content in ANNEXES IV and VII, and concordant provisions, in the latter cases, where appropriate, by means of the Delegated Acts that could be adopted by the European Commission referred to in Article 97 AIA (in relation to Articles 11 and 43 AIA), could be considered suitable.

9. Limitations on collaboration by the initial regulatory space

Most of the provisions that serve to facilitate compliance with the GDPR refer to or assume that we are dealing with high-risk AI systems, including in some cases general-purpose AI models.

While it is true that the existence of high-risk AI systems makes it necessary to apply the GDPR in accordance with the risk assumed for the right to data protection, it is also true that AI systems, even if they are not high-risk, require assuming a regulatory compliance framework from the GDPR perspective, in the terms previously indicated. This would be required for these AI systems not only in relation to compliance with processing requirements and principles, but also in the adoption of technical and organisational measures for the protection of personal data by design and by default (Art. 25 GDPR), with particular importance of the PIA.

This is without prejudice to the adoption of the codes of conduct and governance arrangements referred to in Article 95 AIA, for the voluntary application to AI systems other than high-risk AI systems, of some or all

[100] *"The right to privacy and the protection of personal data must be ensured throughout the entire lifecycle of the AI system. In this regard, the principles of data minimisation and data protection by design and by default, as set out in Union data protection law, are applicable when personal data are processed. Measures taken by providers to ensure compliance with these principles may include not only anonymisation and encryption, but also the use of technology that allows algorithms to be carried over to the data and the training of AI systems without requiring transmission between the parties or copying of the raw or structured data, without prejudice to the data governance requirements set out in this Regulation.*

of the requirements set out in Chapter III, Section 2, which would allow the limitation to be fully or partially overcome.

IV. Final reflections

The purpose of these lines was to analyse, in the search for a certain legal framework, the eventual interactions that occur between the AIA and the GDPR. As we said, how they relate to and complement each other from the point of view of the legal certainty necessary for the preservation of the right to the protection of personal data, in the context of the essential technological development and its regulation.

These are set out as final reflections, also by way of conclusion:

- The GDPR is fully applicable to the entire life cycle of AI systems and their entire value chain, if in its context and to whatever extent personal data are processed or, based on them, the data subject is affected by automated decisions, and compliance, also with regard to the GDPR, is one of the foundations of an ethical AI.

- In cases where the AIA provides for a rule as *lex specialis* (Section III.3) cannot be considered as providing a legitimate basis for the processing of personal data pursuant to Article 8 LED (Recital 23 AIA), the application of the GDPR must be ensured, and the principles of necessity, proportionality, purpose limitation, among others, and the intervention of an independent authority must be ensured.

- In cases of establishment of a legitimate basis for processing, or in the event of lifting of the prohibition of processing laid down in Art. 9.1 GDPR (Arts. 10.5. and 54 AIA), the rules laid down to safeguard the interests and fundamental rights of data subjects must be complied with.

- The intervention of data protection authorities should take place in an area of '*structured and institutionalised cooperation*' with other competent authorities, such as market surveillance authorities. The rules providing for the relationship between the two in the AIA should always be interpreted in the sense most consistent with their independence and their framework for action under the GDPR.

- The guidance of data protection authorities in the area of *sandboxes* need to be formalised in a context of legal certainty, with a complete and precise regulation also of their compliance framework.

- If we analyse the requirements of high-risk AI systems (Chapter III, Section 2, AIA) in particular the human oversight referred to in Article 14(4) and (5), we can conclude that the factual scenario of Article 22 GDPR, in particular in cases where the automated individual decision would be possible because it falls under one of the cases of its paragraph 2, could not take place.

- Regarding the space established around the provision of harmonised standards, common specifications, conformity assessment and certificates referred to in Section 5 of Chapter III, it is clear that insofar as they are aimed at verifying compliance with the provisions of Section 2, they can facilitate the implementation of the GDPR, especially as they will facilitate a standardised framework of interaction between the actors involved, and even the identification of controllers and co-controllers of personal data processing in the value chain.

- In this context, the principles, requirements and measures relating to the processing of personal data (GDPR) contained, as a statement, in the above-mentioned Recital (69) and reproduced above, could be given substance and made possible for verification, where appropriate by amending ANNEXES V, paragraph 5 (concerning the declaration of the existence of processing of personal data and compliance with the GDPR), IV ("*Technical documentation referred to in Article 11(1)*") and VII ("*Conformity based on the assessment of the quality management system and the assessment of the technical documentation*"), where appropriate by means of Delegated Acts that may be adopted by the European Commission pursuant to Article 97 AIA (in conjunction with Articles 11 and 43 AIA).

- The adoption of the codes of conduct and governance mechanisms referred to in Article 95 CPR should be promoted for the voluntary application to AI systems other than high-risk AI systems of the provisions of Chapter III, Section 2, allowing to contribute to the framework of legal certainty in the application of the GDPR.

- To the extent that there may be difficulties in identifying data controllers in the context of the AI system's lifecycle and *value chain*, and being unable to waive compliance with the GDPR, all actors involved in it must exercise special diligence, adopting the necessary measures, including contractual measures, or even deciding not to integrate the service, in order to ensure compliance with the GDPR. The risk of integrating a non-compliant, unaccountable or undocumented element into the system is thus reduced.

This solution approaches the obligation to declare compliance with the GDPR, as part of the conformity assessment (ANNEX V, paragraph 4a, AIA) when the high-risk AI system referred to involves the processing of personal data.

Artificial Intelligence prohibited or unacceptable for the Regulation (Article 5)

BIOMETRIC RECOGNITION IN THE ARTIFICIAL INTELLIGENCE ACT: EXEMPTIONS, PROHIBITIONS AND HIGH-RISK SPECIALTIES

Leire Escajedo San-Epifanio[1]
Senior Lecturer in Constitutional Law
at the University of the Basque Country/ Euskal Herriko Unibertsitatea

I. Biometric recognition in the Artificial Intelligence Act: key aspects of its regulation

1. The regulatory approach to biometric recognition in the Regulation

In the last two decades, the terms *biometrics* or *biometric recognition* have become associated with automated recognition systems that are based on anatomical-physical, physiological or behavioural characteristics of individuals[2]. This recent association, together with the fact that definitions of biometric data often contain examples such as fingerprint, face or iris scanning, explains that when dealing with automated biometric recognition, very relevant aspects of the discipline of biometrics, its uses and its state of the art tend to be overlooked. Etymologically, the word *Biometrics* comes from the Greek words *bios* (life) and *metron* (measure)[3], and its first formal definition, but not the first use of the term, is attributed to Francis Galton, co-founder in 1901 of the journal *Biometrika*[4].

[1] This work is part of the coordinated project *AI-Biosurv – Biosurveillance through AI in the Post-Covid Era. Corporeidad, Identidad y Ciberseguridad* (MICINN, Proyectos de Transición Ecológica y digital, TED2021-129975B-C21), PI L., Leire Escajedo San-Epifanio; and of two Infrastructures, *Enhanced Experimental Infrastructure for 5G and 6G Research and advanced services SN4E+* (Enhanced Smart Networks for Everything), and Experimental Infrastructure for 5G and 6G Research and advanced services *SmartNets4E* (Smart Networks for Everything), IP E.Jacob Taquet.

[2] See Busch, C., „Biometrische Verfahren – Chancen, Stolpersteine und Perspectiven", in P. Schaar (ed.), cit., 2007, 29; Lassman, G., *Bewertungskirterien zur Vergleichbarkeit Biometrischer Verfahren*, TeleTrust Deutschand, 2002.

[3] Escajedo San-Epifanio, L., *Biometric Technologies, Identity and Fundamental Rights*, Thomson Reuters Aranzadi, 2017, 44-45.

[4] Among the first uses of the term is Christoph Bernoulli, who in 1841 used the term *biometry* to refer to taking measurements of human beings for statistical purposes. See Saborowksi, M., „Die Pluripotenz der Biodaten. Beobachtungen zu einem Verwertungsgeschehen",

The term encompasses a very broad set of methods that allow for the measurable study of all types of biological phenomena or processes that occur in living organisms – whether human or non-human[5]. The International Biometric Society, founded in 1947 and present in more than 60 countries, describes itself as promoting "the *development and application of mathematical and statistical theory and methods to the biosciences, including agriculture, biomedical sciences and public health, ecology, environmental forestry and* related *disciplines*"[6]. This includes older methods, of course, than Artificial Intelligence or computation, and even the voice of *Biometrics.* It has even been said that the encounter between Life Sciences and metrics represents an immense chapter in the History of Science[7].

From this broad reference, and insofar as it refers to the human being, in a first approach, the expression can be used to refer to *any data* obtained from biological properties, behavioural aspects, physiological characteristics, habits, or actions that have been obtained by means of some kind of measuring method or technique[8]. And it is precisely this last detail, the measurative processing, that, over and above its link to corporeality, characterises biometric data. Compared to other datasets obtained from a person's body, biometric data are not so much distinguished by accurately describing "natural properties" -or attributes-[9] , but by being an expression of the measurative processing of these[10] .

Until the recent adoption of the AIA, the biometric data that had received most legal attention were those described in Article 4.14 of the General Data Protection Regulation (GDPR). Indeed, some texts in the legal literature tend to consider that only '*personal data obtained from a specific technical processing, relating to the physical, physiological or behavioural characteristics of a natural person which allow or confirm the unique identification of that person [...]*[11]' (Art. 4.14 GDPR) are

in Potthast, T./ Herrmann, B./ Müller, U. (eds.), *Wem gehört der menschliche Körper?*, Mentis, Paderborn, 2010, 380.

[5] Escajedo San-Epifanio, L., *Biometric Technologies*, cit. 2017, 27-28.

[6] *Ibid.*

[7] Albrizio, A., „Biometry and Antrophometry", *Journal of Anthropological Sciences*, vol. 85, 2007,101-123, 102-106; Abs, M., „Biometrik", 1971,945-946; Ghilardi, G./ Keller, F., „Epistemological Foundations of Biometrics", in *Second Generation*,24-25.

[8] This notion is discussed below, at II.1.

[9] Saborowksi, M., „Die Pluripotenz der Biodaten", cit., 2010,367-368.

[10] Mordini, E./ Tzovaras, D./ Ashton, H., in Mordini, E./ Tzsovaras, D. (eds.), *Second Generation Biometrics: The Ethical, Legal and Social Context*, Springer, 2012,7-8.

[11] The reference to Article 4.14 GDPR, in which facial images and fingerprint data are mentioned as examples of biometric data, is deliberately excluded here. This will be returned to in II.2.

biometric data. The fact is, however, that not all biometric data have this potential to enable or confirm the unique identification of a person[12] .

The notion of art. 4.14 GDPR is consistent with a time when identity recognition technologies, in particular biometric authentication, dominated the landscape of technologies implemented in real life. Today, however, there are many recognition technologies that offer non-individualising recognition utilities, e.g., for the purpose of determining a person's age, health status or stress level, or to pinpoint the emotions they are going through. As these latter systems do not have the potential to serve as a basis for a unique identification, they may fall under the concept of personal data, but not, strictly speaking, under the notion of Art. 4.14 GDPR[13].

In the absence of a clear legal status for these non-identifying biometric data, legislators found it necessary to tackle this task in the process of drafting the AIA, which, along with biometric identification utilities, will refer to emotion recognition and categorisation. The Commission[14], the Parliament[15], and the Council[16] offered different alternatives to complete the limited notion of art. 4.14 GDPR for the purposes of the AIA, but, as will be seen in section II.2, at some point in the trialogues a new proposal ended up being imposed.

In addition to clarifying this notion of biometric data, legislators had to answer other questions relating to biometric recognition systems that did not have adequate regulatory precedents. Without seeking to be exhaustive, choosing the most relevant ones for the purposes of this exposition, legislators had to decide whether AIA would encompass all conceivable automated biometric recognition (from verification and identification, to emotion rec-

[12] Kindt, E., "Having yes, using no? About the new legal regime for biometric data", *Computer Law & Security Review,* 34 (3), 2018, 523-538.

[13] Concerning the processing of special categories of data.

[14] The main reference used for the interpretation of the Commission's position, unless otherwise indicated, has been the *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence*, {SEC(2021) 167 final} -{SWD(2021) 84 final}- {SWD(2021) 85 final}.

[15] The *Amendments adopted by the European Parliament on 14 June 2023* on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union (COM(2021)0206 -C9-0146/2021- 2021/0106(COD)) have been used as a reference for the position of the European Parliament.

[16] The *compromise text of the Fourth Presidency* on the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union, Brussels, 19 October 2022, was used as the main text for the Council's position. Interinstitutional file 2021/0106 (COD).

ognition, to broad categorisations and screening) or whether such a holistic ambition would be discarded.

Secondly, given that the AIA is essentially a pre-market risk assessment regulation, there was a question as to which types of biometric recognition would in any case be excluded from market access and which would be treated as high-risk modalities.

Finally, thirdly, it remained to be clarified what would happen to biometric recognition systems (mostly for national identity management or border identification) which, before the entry into force of the AIA -which has not yet happened at the time of writing- were already operated by Member States in their territories or even at Community level as part of the *area of freedom, security and justice* (such as the SIS or EURODAC system)[17].

2 Relevant background: the gap between the White Paper and EP Resolutions on some modalities of biometric recognition

As far as biometric recognition is concerned, the gap between the Commission's White Paper on Artificial Intelligence (2020)[18] and some previous Parliament positions on biometric recognition, including two 2021 resolutions, made the search for agreement in the AIA process a difficult one.

The Commission, in the White Paper, identified remote biometric identification as an example of AI applications which, independently of the sector in which they are applied, could be considered as "high risk"[19]. Consistently, its proposed Regulation will capture most remote biometric identification systems in high-risk modalities. It should not be misleading that the same proposal includes the prohibition of the "use" -but not the prohibition of placing on the market- of some forms of remote biometric identification among the practices to be prohibited in its proposed Article 5(1)(d). The large number of cases that could be covered by the exceptions that the Commission admitted with respect to this prohibition practically emptied the latter of its content[20].

This will create an important paradox, because the GDPR prohibits with few exceptions the processing of biometric identifiers -essential in remote biometric identification systems- and the AIA proposal seems to open up

[17] See infra, VI.

[18] European Commission*, White Paper on Artificial Intelligence – A European approach to excellence and trust*, Brussels, 19.2.2020, COM (2020) 65 final.

[19] European Commission, *White Paper*, cit., 2020,22.

[20] See below, III and IV, on recognition systems falling under the prohibited practices of Art. 5.

the 'possibility of use' of these data as high-risk modalities. The confusion increases if we take into account that the Commission, in the White Paper -and later in the Explanatory Memorandum of the proposal- will recall that, in accordance with existing data protection rules and *the EU Charter of Fundamental Rights*, AI can only be used for remote biometric identification purposes when such use is duly justified, proportionate, and subject to *appropriate* safeguards[21]. In view of such recognition, it seems inappropriate that AIA and GDPR, while respecting the purpose of each of these rules, are not adequately coordinated in their treatment of the 'use' of remote biometric identification systems.

In any case, and in line with its position on AIA, the Commission's efforts with regard to biometric recognition technologies in general and prohibited modalities in particular, will focus on classifying them by level of risk, clearly delimiting what the cases of justified and proportionate use may be, and setting out some of the applicable safeguards. It is also appropriate to note that the White Paper will propose a distinction between remote biometric identification on the one hand, and authentication on the other hand. This is unusual because the GDPR, it should be noted, already prohibited the use of biometric data for the purpose of unique identification, except in very specific cases (Art. 9.1 and 9.2. GDPR), regardless of whether they were used for verification or *one-to-one* (1-to-1) or *one-to-many (*1-to-n) identification functionalities. The Commission, however, only states in the White Paper that identification utilities will be treated differently from identification utilities, without specifying further. The distinction will prove to be very relevant given that, in the final text of the AIA, as will be seen, biometric verifications -interpreted in a very extensive way- will end up being excluded even from the high-risk modalities[22], leaving them at most to the discretion of possible voluntary Codes of Conduct[23].

As far as the European Parliament is concerned, it published two important resolutions prior to the AIA in which a broad rejection of most forms of biometric recognition, in particular when used by law enforcement authorities, was expressed.

In the European Parliament Resolution of 6th October 2021[24] on the use

[21] European Commission, *White Paper*, cit., 2020,26-27.

[22] Vid. infra. II.3. Biometric verification.

[23] Vid. infra. V. High-risk modalities.

[24] European Parliament resolution of 6 October 2021 *on Artificial Intelligence in criminal law and its use by law enforcement authorities in criminal matters* (2020/2016(INI)), published in Official Journal C 132/17 of 23 March 2022.

of AI by law enforcement authorities in criminal matters. The Parliament considers, firstly, that the deployment of facial recognition systems should be limited to clearly justified purposes and be done in full respect of the principles of proportionality and necessity and of the applicable law. Secondly, it calls for a permanent ban on the use of automated analysis or recognition in publicly accessible spaces of human characteristics such as gait, fingerprints, DNA, voice and other biometric and behavioural signals. And finally, thirdly, it calls for a moratorium on the use of facial recognition until the following circumstances are met: that the technical standards can be considered fully in line with fundamental rights; that the results are not biased or discriminatory; that the regulatory framework is strict; and that there is empirical evidence of the necessity and proportionality of the deployment of these technologies, with the sole exception of the case where they are used strictly for the identification victims of crime[25].

It should also be noted that before that date the European Parliament had already recommended a ban on the use of automated biometric recognition applications such as facial recognition for educational and cultural purposes, in particular with regard to minors, unless their use was expressly authorised by law[26]. This call for restrictions and moratoria on automated biometric recognition will characterise the Parliament's position towards the Council until the last gasp of the adoption of the AIA.

3. Regulatory scheme of biometric recognition technologies in the Regulation

Automated biometric recognition is covered throughout the AIA in a significant number of recitals, definition sections and substantive articles. In the latter set, five different legal statutes, applicable to certain sets of biometric recognition technologies, stand out in particular.

Firstly, we find a series of biometric practices prohibited in the only article of Chapter II (Art. 5. AIA). A second statute is that of biometric practices considered to be high-risk (Art. 6 et seq., supplemented by Annex III). Thirdly, a number of biometric practices should be noted which, as a result of the provisions of Art. 2.3, are excluded from the scope of application of the AIA. Fourthly, on the basis of Article 111 AIA with the addition of Annex X, a specific statute is foreseen for a set of biometric recognition practices which

[25] Rostalski, F./ Weiss, E., „Verbotene KI-Praktiken", in Hilgendorf, E./ Roth-Isigkeit, D. (eds), *Die neue Verordnung der EU*, cit., 2023, 47-48.

[26] Point 45 of the European Parliament resolution of 19 May 2021 on Artificial Intelligence in the educational, cultural and audiovisual sectors (2020/2017(INI)).

are used in the field of large-scale IT systems established by EU legislation in law enforcement and border control matters. Finally, and fifthly, a systematic interpretation of the AIA brings to light a set of biometric recognition systems which, because of the little or no attention they receive in the AIA, seem to be outside its scope or at least in doubt.

With regard to the first group of practices, that of the biometric recognition modalities affected by the AIA prohibitions, six groups of practices are listed in Article 5[27]:

1. certain biometric systems that can be used to *evaluate or classify* natural persons or groups of persons *on the basis of their social behaviour or* known, inferred or predicted *personal or personality characteristics* (Art. 5.1.c);

2. some of the biometric systems which, by means of *profiling or personality assessment*, can be used for risk assessment of natural persons *in order to assess or predict the risk of committing crimes* (Art.5.1. d);

3. certain biometric recognition systems that may be used in the *creation or expansion of certain* facial recognition *databases* (Art.5.1. e);

4. certain recognition systems that make it possible to *infer the emotions* of a natural person *in workplaces and educational institutions* (art. 5.1. f);

5. some biometric categorisation systems that individually classify natural persons on the basis of their biometric data in order to *deduce or infer their race, political opinions, trade union membership, religious or philosophical convictions, sex life or sexual orientation* (Art. 5.1. g).

and 6. some *"real-time" remote biometric identification* in publicly accessible areas for law enforcement purposes.

A second set of biometric technologies, within the meaning of Article 6 and in particular Annex III of the Regulation, is classified as *high-risk systems*[28]. This set comprises, as a first sub-group, remote biometric identification systems, biometric categorisation systems based on sensitive attributes or characteristics and certain emotion recognition systems, all of them irrespective of the operational scenario in which they are applied, provided they are not among the prohibited categories. Excluded are, in any case: recognition systems covered by the definitions of prohibited AI; systems used by Member States for military, defence or national security purposes (Art. 2.3 AIA); and, as indicated in Annex III itself, recognition systems providing authentication or verification functionalities.

The second sub-group of high-risk biometric recognition is that of the systems that can be considered to be included in sections 2 to 8 of Annex

[27] See below, III and IV.

[28] See infra, at V.

III AIA. These sections of the Annex provide a list of 21 high-risk forms of AI, grouped into six operational scenarios: education and vocational training; employment; essential services and benefits; law enforcement; cross-border transit; and administration of justice[29]. The wording of these high-risk AI modalities is strongly formulated as "*AI systems intended to be used for*" actions such as assessing (risks, outcomes, learning levels, reliability), tracking, detecting prohibited behaviour, classification or decision making, and, as will be detailed below, may potentially cover some biometric recognition functionalities not listed in Annex III.1.

From a legislative perspective, this wording generates legal uncertainty and, to some extent, apparently contradicts the horizontal, risk-focused regulatory model of the AIA. Thus, in relation to this last issue, it should be noted that: some of the provisions of these lists, organised by operational scenarios, overlap with the subgroup of section 1 of Annex III, making reiteration unnecessary; and in other provisions, for their part, there is an overlap with the necessary assessment of proportionality that the GDPR establishes with regard to identifiable data.

The third important set of biometric technologies is affected by the provisions of Articles 111 et seq. of the AIA, in connection with Annex X. These are large-scale IT systems that have been put into service or will be put into service within 36 months of the date foreseen for the entry into force of the AIA)[30]. As will be detailed in Section VI, these are large-scale recognition systems regulated by EU legislative acts that, moreover, are already in use or in the process of being implemented for the management of the areas of freedom, security, and justice. The plan is that, with the exception of the application of the prohibitions provided for in Article 5.1 of the AIA (the form in which they are to be applied has yet to be determined), these large-scale systems will enjoy a temporary moratorium on the application of the AIA, which, as will be seen, will be extended until practically January 2031. After this period, moreover, the fact that other provisions of the AIA potentially leave some of the verification and non-remote identification utilities offered by these systems outside its scope of application will have to be addressed.

In fifth and final place, as noted above, a systematic interpretation of the AIA brings to light a fifth statute applicable to certain forms of biometric recognition: those excluded from the AIA. In addition to the aforementioned

[29] For details on high-risk systems, see the commentary on Articles 6 and following by L. Cotino Hueso in this work.

[30] The final provisions state that the Regulation will be fully applicable three years after its publication in the Official Journal, which at the time of writing has not yet taken place.

situation of verification and non-remote identification, which will be dealt with in Section II, the final text of art. 2.3 AIA, concerning the scope of application, indicates that the AIA does not apply to AI systems that, and to the extent that they are placed on the market, put into service or used, with or without modification, exclusively for military defence, or national security purposes, regardless of the type of entity carrying out these activities. These systems are, however, subject to the GDPR and its derived acts.

II. Some key concepts in the typification of biometric recognition modalities in the statement of the Regulation: biometric data and biometric verification

Article 3 of the AIA contains a long list of definitions. Some of them, such as remote biometric identification (in real time and delayed), or categorisation and profiling, will be the object of attention as part of the typification of factual assumptions offered by the AIA. However, there are other notions that have a transversal relevance in the set of articles that refer to biometric recognition, and which, for this reason, deserve to receive attention in these first sections. This is the case of the notion of "biometric data" used by the AIA -departing from Art. 4.14 RGPD – and of the notion of "biometric verification", as a category that the AIA insistently tries to distance from the concept of biometric identification, in particular from the remote modality.

1. A new notion of "biometric data" to clearly encompass non-differentiating biometrics?

1.1. The situation prior to the adoption of the Regulation: the GDPR notion of biometric data versus the scientific-technical notion

It was noted in the introduction that, until the adoption of the AIA, the most relevant legal notion of "biometric data" was that contained in Article 4.14 of the GDPR. The last paragraph of Art. 4.14 GDPR will be excluded from our reflection, because it is somewhat unfortunate. In it, the legislators present facial images and dactyloscopic data as examples of biometric data[31].

[31] Sumer, B. "When do the images of biometric characteristics qualify as special categories of data under the GDPR: a systematic approach to biometric data procesisng", BIOSIG 2922 -International Conference of the Biometrics Special Interest Group, published in open access in IEE Xplore; Romeo Casabona, C., "Biometric data (Commentary on Article 4.13 RGPD)", in *Comentario al Reglamento General de Protección de Datos y a la Ley Orgánica de Protección de Datos personales y Garantía de los Derechos Digitales*, A. Troncoso Reigada (dir.), Vol. 1, 2021, 709-714.

However, neither a photo of a face nor fingerprint information -including a fingerprint print[32]- are in themselves biometric data. They are, admittedly, possible sources of biometric data, but it will be a measurable processing that will determine whether or not unique biometric data are obtained from them[33].

Excluding, therefore, that subparagraph, it can be said that, according to the GDPR, biometric data are "*personal data obtained from a specific technical process relating to the physical, physiological or behavioural characteristics of a natural person which allow or confirm the unique identification of that person*" (Art. 4.14 GDPR). A comparison should now be made between this notion, on the basis of which the GDPR establishes a category of specially protected data, and the definition of biometric data consistent with the entity of the discipline of biometrics, presented in section I.1 of this paper.

Even before the adoption of the GDPR, different experts were already warning about the gap between, on the one hand, the notions of biometric data contained in different legal documents, whether binding or not, and, on the other hand, the concept of biometric data in a scientific perspective[34]. As Kindt and Jasserand pointed out, there was a significant misalignment between the technological possibilities and the legal definition[35], a problem that was exacerbated with the adoption of the GDPR.

The following definition of the scientific-technical notion of biometric data is proposed here, postulating that it is desirable that it be progressively taken into account in the design of regulatory frameworks for biometric recognition techniques[36]:

> *Biometric data are those obtained from people's bodies as an expression of some kind of measurement study and, in its vast set, it is relevant to distinguish between two large groups, depending on whether the data are obtained from static biometrics or dynamic biometrics*[37].

[32] See on this concept of a biometric data source below, II.2.

[33] Kindt, E., Having yes, using no? About the new legal regime for biometric data, *Computer Law & Security Review*, Volume 34, Issue 3, 2018, 523-5388.

[34] Jasserand, C. A., "Avoiding Terminological Confusion between the Notions of "biometrics" and "biometric Data": An Investigation into the Meanings of the Terms from a European Data Protection and a Scientific Perspective", *International Data Privacy Law* 6 (1), 2015.

[35] Jasserand, C. A., "Avoiding Terminological Confusion", cit., 2015; Kindt, E., "Having yes, using no?", cit., 2018, 523-538.

[36] Kindt, for his part, proposes to define biometric data as "*all personal data that (a) relate directly or indirectly to unique or distinctive biological or behavioural characteristics of human beings and (b) are used or suitable for use by automated means (c) for the purpose of identification, verification of identity or verification of a claim of a living natural person*", although by "personal" he means data from individuals and not, therefore, data that fit the GDPR definition of personal data. Kindt, E. *Privacy and Data Protection Issues of Biometric Applications -A Comparative Legal Analysis*, Springer, 2013, 11.

[37] Escajedo San-Epifanio, L., *Biometric Technologies*, cit., 2017,100-101.

Static biometrics merge together those methods that capture specific metric information from the anatomical-physical characteristics of a human body. Dynamic biometrics, on the other hand, comprise those methods that are applied to capture sequential or cyclical information from a human body on motor skills, as well as body signs and parameters in a broad sense, whether in their external or internal, voluntary or involuntary dimensions.

Note, because the distinction will be relevant, that static and dynamic biometrics differ in two key respects: first, by the type of sources from which they obtain information (anatomical-physical characteristics versus, in some way, the body in operation); and secondly, in the fact that the capture of static biometrics can be done in an instant, while the capture of dynamic biometrics requires a longer or shorter period of time. In other words, biometrics that capture the pattern of a fingerprint or an iris, as well as the geometry of a face, act on a specific raw data point. For example, a fingerprint is extracted from the dermatoglyphs of a finger and this print is sufficient to start the process of biometrising the attributes that distinguish that fingerprint from others. Technologies based on dynamic biometrics, on the other hand, such as those that analyse the spectrum of a person's voice, the pattern of their walking gait or the speed of their heartbeat, are technologies that need to be able to capture information from the source over a longer or shorter period of time.

1.2. The need for a functional notion covering biometric data with and without identifying potential, whether personal or not

According to the Explanatory Memorandum of the AIA Proposal, this regulation complements -without displacing- the framework of guarantees already contained, among others, in legal texts such as the GDPR. With respect to the latter, the Explanatory Memorandum also states that the GDPR is a *lex specialis*, but with a fundamentally complementary character. Despite this statement, however, in the specific case of the use given in the AIA to the notion of biometric data, with its own definition, the difference with respect to the GDPR is very relevant[38]. This difference will be described in this section, leaving for Section II.1.3 both the way in which the Commission, Parliament and Council respectively proposed to deal with this circumstance, and the option that was finally incorporated into the AIA.

Although it repeats the criticised final clause of Art. 4.14 GDPR referred

[38] Czarnocki, J., "Will new definitions of emotion recognition and biometric data hamper the objectives of the proposed AIA?", *International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2021, 1-4.

to in II.1.1. above, the notion of biometric data used in the AIA differs from the one in which the requirement that the data allow or confirm "*the unique identification of a person*" has been removed. Consequently, the set of data that may be used in biometric recognition systems will include data subject to the scope of application of the GDPR insofar as they can be considered personal, some of them also belonging to the category of special protection (Art. 9 GDPR). As far as the AIA is concerned, some dimensions of biometric data processing will be subject to its application -irrespective of whether or not they are unique data and even without the need for them to be personal data-and, at the same time, AI verification systems using special category biometric data (Art. 4.14 and 9 GDPR) will be excluded from the AIA[39]. In terms of safeguards, it has to be said, this system -with the addition of the complicated wording of some articles of the AIA- is rather worrying.

In this context, it is necessary to make a series of notes on the characteristics that biometric recognition systems that aim to offer an identification functionality (or singularisation of identity) must have, as opposed to the characteristics of those systems that capture other types of biometric information that do not singularise:

a) Characteristics of biometric recognition systems with singling-out potential

Recognition systems with unique identification potential can only be based on biometric datasets that meet a number of characteristics[40], including universality, uniqueness potential, inherence or permanence. A biometric recognition system with unique identification potential must be based on a characteristic or attribute that is *universal,* in the sense that the source of biometric information is, in principle, available to all human beings (or with very few exceptions).

In addition to this universality, the source of information must also allow the *presence of singularising elements* to be captured. Fingers are almost universally present and most people have dermatoglyphs, or skin folds, that leave fingerprints when they touch surfaces. A biometric recognition system that quantifies the number of fingers on the right hand is not sufficiently unique due to the high concurrence of the number of five fingers on most human hands. The case of dermatoglyphs, on the other hand, is different. It is a source of information in which we can capture such a high degree of uniqueness that,

[39] On this, see. II.3.

[40] Detailed analysis of these characteristics and of approaches such as the so-called *Seven Pillar sor Biometrics Wisdom*, with detailed references, in Escajedo San-Epifanio, L., *Tecnologías Biétricas*, cit. 2017, 88-98.

statistically, it is considered unlikely that there are two fingers -even in the same person- with an identical fingerprint. This makes fingerprint biometrics with a high uniqueness potential, a potential that can also be captured. This contrasts, for example, with DNA, which, while being highly unique in humans, it cannot currently be captured and processed in an automated way.

Finally, in order to support an identification system, it is also very relevant that the biometric reference information is information *that is inherent* to the body of the person -or at least reasonably difficult to modify or supplant at will- and that it *remains* despite the passage of time[41]. Permanence is not required in absolute terms, but rather in the sense that it has sufficient stability to be useful for re-identifying the person after a reasonable period of time. The pinna pattern, for example, remains even though the size of the ear varies with age. The example of the ID card image, which is not a biometric but a data source, can also be used. This image, because of the requirements under which it is taken, is reasonably estimated to be able to re-identify someone within ten years of its issue or even longer when the person has reached a certain age, hence the issuing of a permanent ID card from the age of 70 onwards.

It should be noted, in this sense, that permanence guarantees that the effort of deploying a biometric identification system in the strict sense will be compensated by the possibility of using it over a long period of time. Regarding to other biometric attributes, the universality and ease of capture contrasts with the limited permanence of the data obtained. Environmental and/or temporal conditions have a significant impact on these data[42]. For example, a person's weight or hair length (unless they have no hair) are easy to measure, but in the same person they are subject to significant variability over the course of a lifetime. It would not make sense to build a unique identification system on the basis of data on these characteristics. Dermatoglyphs, however, finish forming around the third or fourth month of foetal development and are maintained -except external injury- throughout a person's life, and can even be captured at some post-mortem stages. In fact, the scientific discipline of necropapiloscopy has techniques for capturing them that can even be applied to petrified or mummified fingers[43].

[41] European Commission/ DG JRC/ Institute of Prospective Technological Studies (IPTS), *Biometrics at the Frontiers: Assessing the Impact on Society, European Commission,* 2005, p.37; Mordini, E./ Massari, S., "Body, Biometrics and Identity", loc. cit., 2012; ZHANG, D./ LU, G. *3D Biometrics. Systems and Applications,* Springer, 2013, 9-12.

[42] Escajedo San-Epifanio, L., Biometric Technologies, cit., 2017, 93-96.

[43] Alegretti, J. C./ Brandimarti de Pini, N. M., "Necropapiloscopy. Identificación de cadáveres y restos humanos", in *Tratado de papiloscopía*, La Rocca, Buenos Aires, 2007, 245-263.

(b) Data characteristics that do not aim at unique identification, but at capturing information of other usefulness

With regard to non-identifying categorisations and, in particular, emotion recognition, recognition systems that apply this type of modality tend to use *soft* biometrics, with low permanence range and identifying potential[44]. The sources, however, are universal sources and have the capacity to capture attributes in a categorising way, in the sense of associating, on the one hand, the attributes freshly captured regarding a person with, on the other hand, average patterns that have been elaborated as a reference range for each of the categories in which it is intended to classify humans (for example, classification by age ranges, interpretation of basic emotions, or detection of suspicious behaviour).

The case of emotions is undoubtedly one of the simplest to explain. The system has been trained, for example, to recognise in faces a range of eye, eyebrow, forehead and chin movements that, in large sets of people, are associated with a high probability with emotions such as anger, joy or fear. Each time a face is placed in front of the system, it will search it for an emotion, trying to find a reasonable match to one of the predefined biometrically predefined categories available to it. Information of this kind can, where appropriate, be used as a guide to persuade someone to purchase certain products or services.

1.3. Evolution of the notion of biometric data in the process of elaboration of the Regulation

In case the current lack of coordination between the notions of biometric data respectively provided for in the AIA and the RGPD is ever reviewed, it is interesting to note how the notion finally incorporated in the AIA was forged. The Commission, Parliament and Council offered different alternatives, which will be analysed, although, surprisingly, at some point in the trilogues, a notion that claims to be inspired by the GDPR ended up being imposed, but, as has already been mentioned, in reality it is far removed from it.

Recital 7 of the Commission Proposal stated that its notion of 'biometric data' was in line with the notion of 'biometric data' as defined in Article 4(14) of Regulation (EU) 2016/679 of the European Parliament and of the Council[45]; in Article 3(18) of Regulation (EU) 2018/1725 of the Europe-

[44] Escajedo San-Epifanio, L., *Biometric Technologies*, cit., 2017, 105-106.

[45] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of individuals with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation) (OJ L 119, 4.5.2016, p. 1).

an Parliament and of the Council[46]; and in Article 3(13) of Directive (EU) 2016/680 of the European Parliament and of the Council[47]. Consistently, it was indicated that such a notion should be interpreted in line with that of the GDPR. In itself, the notion did not raise doubts. It was problematic, however, that neither this nor other notions addressed the fact that the processing of non-identifying biometric data by means of AI was possible.

The Parliament, on the other hand, did address the latter need. The notion of biometric data that it adopted in its position coincided literally with that of the Commission, but it proposed to add a new concept to the list of definitions in Art. 3 AIA: the concept of *biometric-based data*. This second notion, differentiated from the notion of *biometric data*, is proposed as a way to complement the functional limitations of Art. 4.14 GDPR for the purposes of the AIA. Biometric-based data, according to the amendment of the European Parliament proposing its inclusion as number 33a of Art. 3, are defined as "*data obtained from specific technical processing relating to the physical, physiological or behavioural signals of a* natural *person*"[48].

This second concept is, however, discarded in the latest version of the AIA, it being considered preferable to use a single notion of biometric data for the whole text. In order to reduce legal uncertainty, in any case, the Commission's initial proposal for a concept of personal data, supported by the positions of the Council and the Parliament, is modified. Thus, according to Article 3(34) of the AIA, for the purposes of this text, biometric data are '*personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, such as facial images or dactyloscopic data'*. The recitals say that this notion should be interpreted in the light of the concept of biometric data in Article 4.14 GDPR, but the differences between the two notions are obvious.

[46] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of individuals with regard to the processing of personal data by Union institutions, bodies, offices and agencies, and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC (OJ L 295, 21.11.2018, p. 39).

[47] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data and repealing Council Framework Decision 2008/977/JHA (Criminal Data Protection Directive) (OJ L 119, 4.5.2016, p. 89).

[48] Amendments adopted by the European Parliament on 14 June 2023, cit. supra.

2. Common minimum of all biometric recognition systems

2.1. Biometric recognition systems

Whether identifiable or not, in order for a biometric data to serve as the basis for a biometric recognition system, a system must be built. Specifically, a system capable of examining a person's body and "recognising" in it, either occasionally or sequentially, some attribute that can be linked to information that, as a reference, has been previously stored -in a database or in external storage devices, such as the chips that some identification documents carry. There were non-automated biometric recognition systems, such as the legendary *Bertillonage*, which will be briefly referred to *below*, but nowadays the term biometric recognition system is commonly applied to structures that combine *hardware* and *software,* increasingly tending to be AI-driven.

In their basic architecture, automated biometric recognition systems need a minimum of modules that usually operate sequentially[49]: interfaces incorporating readers or sensors, modules for enhancing capture and attribute extraction, storage spaces and, of course, modules for matching freshly captured information against stored information.

On the other hand, biometric recognition systems comprise a series of processes or phases that can generically be grouped into four phases[50]: a) a recruitment phase, either of a person in the singular, or of a category such as "emotion fear", "age group over 65", or gender categories; b) a phase of elaboration of the pattern (which can comprise several sub-phases); c) a third moment of capturing the fresh pattern, that is, a moment in which we place a person in front of the system, so that he/she recognises in that body information that can be captured and matched; and, d) finally, the matching of the freshly captured pattern with the stored pattern.

2.2. Data sources, raw data and biometric standards

In absolute terms, there is no single biometric that can be considered the best for recognition. What is to be recognised on a body can be very diverse, and the possibilities of the operational scenarios in which the recognition is to be applied must also be taken into account. For example, fingerprints are very unique, but at the entrance and exit of an archaeological excavation or construction site, or at the entrance and exit of a school canteen, it is unlikely that the average person's hands will be clean enough to be able to recognise them well. If the group of enrolled subjects is not very large, a recognition

[49] Escajedo San-Epifanio, L., *Biometric Technologies*, cit., 2017, 177-178.

[50] *Ibid*, pp. 178-180.

system based on hand geometry would be of better service, although in net terms it is less unique.

From a scientific-technical point of view, an important distinction is made between three concepts. As a first concept, *biometric data sources* are those anatomical-physical or dynamic characteristics of the human body from which information is captured for the operation of a biometric recognition system, regardless of whether the biometric recognition system is identifiable or not. Secondly, *raw data* are the biometrisable attributes that are present in that source and can be captured as such. Thirdly and finally, *biometric templates* (*empreinte numérique*) are the transformed or technified versions of the captured information. These patterns are susceptible to digital storage and subsequent comparison and, strictly speaking, coincide with the notions of biometric data offered by Art. 3.34 AIA and Art. 4.14 RGPD, in the sense that they are the result of a technical processing of physical, physiological or behavioural characteristics of a natural person.

Each recognition system, in addition to being based on a particular source of information, is designed to process in a particular way the unique attributes present in the raw data. Thus, for example, dermatoglyphs, or the patterns that skin folds, ridges and dermal grooves form on the palms of the hands, soles of the feet and the fingertips, have been considered for two centuries as sources of highly unique data[51]; statistically, the possibility of two fingers -even from the same person- having an identical fingerprint is thought to be bordering on the impossible. However, to obtain a biometric pattern, we need to capture some kind of image of the fingerprint, apply a series of filters to it, and from there, each recognition system will proceed to capture biometric information. Some fingerprint recognition systems, for example, proceed to pinpoint the position or geolocation on the print of a number of points or minutiae that they consider most striking. The processing will not store the particularities present in those points (e.g. some striking bifurcation), but only the position(s) where within that fingerprint some particularity is present. This information regarding the location of unique characters is not as unique as a complete matching of the prints, but statistically it is considered unique enough to serve as a basis for an identification system.

It should be noted that a biometric pattern never stores the totality of characteristic attributes that may be present in a raw data. It can be said that such a pattern is always intended to be the best possible synthesis of what is characteristic of a raw data. This is, however, an intention, and not all recognition systems are equally reliable. If the system is not of high quality and

[51] *Ibid.* pp. 109-111.

there are many people enrolled, there is an increased risk that two practically identical digital patterns for two different dermatoglyphs will appear on the source[52] or that the system will not be able to positively recognise a match between the raw data of an enrolled subject and the data that was stored during enrolment.

2.3. Biometric utilities that identify, utilities that don't

The existence of prior individualised enrolment is an element that, in the light of what has been described in the previous sections, clearly distinguishes systems that offer unique identification utilities from those that do not.

The identification utilities offered by biometric recognition systems in a broad sense can be grouped into three main groups[53]: verification utilities (or verification of claimed identity, one-to-one); identification utilities in the strict sense (or determination of identity without prior claim of identity, one-to-many); and *biometric screening* utilities or utilities for searching and locating certain unique individuals in unbounded environments -physical or virtual- by searching, in some way, for a subject x, i.e., not enrolled, in an infinite mass.

From a legal point of view, these identification utilities, especially the first two, tend to be defined separately from others -such as categorisation- because they involve the processing of biometric data that certainly fit the notion of Article 4.14 GDPR. These are data that '*enable or confirm the unique identification of that person*'. In order to offer such functionality, this type of system requires the prior biometric enrolment of the person to be re-identified, in the sense that at a time prior to the identification or verification, the person's information was captured and stored. This storage can take different forms: 1) in some cases there is a reference database that stores the enrolled subjects' information, either as a dedicated database that operates as part of the particular computer system, or as a database that is accessed online; 2) in other cases, such as the biometric passport, the biometric information is stored on a chip inserted in the document, so that, anywhere in the world, and provided that there is a device that combines mechanical reading of the passport and fresh capture of the person's features, it is possible to proceed to verify whether a biometric passport belongs -or not- to the body that carries it.

Categorisation, profiling or emotion recognition, among others, do not aim at unique identification. What they aim to do is to recognise biometric attributes on a person's body -either in its static or dynamic dimension- that

[52] *Ibid.* pp. 109-111.

[53] Wayman, J./ Jain, A.K./ Maio, D., "An Introduction", 2005, 4-5.

have previously been associated with a series of categories. Thus, for example, if in order to distinguish emotions, emotions have been associated with certain movement stripes of the chin or the arching of the eyebrows, the system will try to link the attributes that it fresh captures on a face with one of the available categories.

Historians point to the pioneering contribution of Alphonse Bertillon (his Bertillonage or *signalement anthropométrique*) as the milestone from which a branch of biometrics specifically aimed at the unequivocal identification of individuals began to develop. On 31st August 1832, the fire branding of convicted criminals was abolished in France and, from a practical point of view, determining in which cases an offender deserved an aggravated sanction, as a recidivist, became more complex. The only reference was documentary records -paper entries of more than 5 million people- and they were of no use when the subjects concealed their identity or falsified it, with the intention of avoiding the aggravating circumstance of recidivism[54].

This and other circumstances led to the emergence of a specialised body within the police: the so-called technical police[55]. Joining the Prefecture of the Paris Police in 1879, Alphonse Bertillon took on the task of drawing up a series of "signalétique" (or *signage in English*) files[56] on certain individuals, based on the anthropobiological work of experts such as Lambert Adolphe Quetelet (1796-1874), or Paul Broca (1825-1880) and his collaborators at the Anthropology Laboratory of the *École Pratique des Hautes Études* in Paris. The "signalement anthropométrique" offered a novel system for tracing the documentary records, which were no longer arranged alphabetically but on the basis of detailed metric descriptions of the bodies (such as wingspan, size of the skull or elbow, feet or ears)[57]. Thus, taking the measurements of a repeat offender made it potentially possible to locate his file even if he had not provided his real name.

The development of digital signal processing techniques (DSP – *Digital Signal Processing*) and their projections in possible unique recognition systems

[54] Auger, D., *Biométrie: l'équilibre entre «liberté individuelle» et promesse sécuritaire serait-il impossible?*, 2005, 26-27.

[55] Dias, C., *La police technique et scientifique*, PUF, Paris, 2000, 12.

[56] Madureira, N., "Police without science: criminal investigation in Portugal: 1880-1936", *Política e Sociedade,* 2005, Vol. 42, n.º 3,45-62. Bertillon, A., *Signaletic instructions including the theory and practice of anthropometrical identification*, Werner, Chicago, 1896.

[57] McCarthy, P., "Biometric Technologies", in *Encyclopedia of applied ethics*, 2nd edition, 2102, Elsevier; see also Sutrop. M./ Lass-Mikko, K., "From Identity Verification to Behavior Prediction", *Review of Policy Research*, vol. 29, no. 1, 21 ff.

through voice[58] or fingerprints[59], led to the recognition already in the early 1960s of the important potential of these technologies in order to guarantee high levels of security in access control, use of personal passwords or financial transactions[60]. With the development and implementation of the first systems, the *first generation* of automated recognition systems using biometrics, developed mainly on the basis of static biometric data sources and applicable to small groups of people, was already in its infancy. In the 1970s, recognition systems based on manual geometry[61] were developed and implemented, and the systems were also tested on larger groups[62], in an attempt to improve the speed and efficiency of recognition methods.

The literature considers that it was from that time onwards that a growing interest in the possible governmental uses of automated identification technologies became evident. In the 1980s, retinal biometrics and dynamic signature recognition systems were developed, followed by facial recognition systems. Technologies based on the iris pattern were proposed in the mid-1980s, but did not become a reality until an algorithm reliable enough to capture the uniqueness of this human feature was developed[63].

At that time, precisely because of their level of invasiveness, a progressive expansion of the use of biometric recognition technologies to society as a whole was not foreseen, but this situation changed radically with the 9/11

[58] Pruzansky, S., "Pattern-matching", *Journal of the Acoustical Society of America,* 1963, 35, 354-358; Li, K. P./ Dammann, J. E./ Chapman, W.D., "Experimental studies in speaker verification", *J. Acoust. Soc. Am.*, 1966, 40, 966-978; Luck, J., "Automatic speaker verification using spectral measurements", *J. Acoust. Soc. Am.*, 1969, 46, 1026-1031; Stevens, K./ Williams, C./ Carbonell, J./ Woods, B., "Speaker authentication and identification: a comparison of spectrographic and auditory presentation of speech material", *J. Acoust. Soc. Am.*, 1968, 44, 596-607; Atal, B., "Automatic recognition of", *Proc. IEEE*, 1976, 64(4), 460-474; Rosenberg, A., "Automatic speaker recognition", *Proc. IEEE*, 1976, 64(4), 475-487.

[59] Trauring, M., "Automatic comparison of finger-ridge patterns", *Nature*, 1963, 197, 938-940.

[60] Trauring, M., "On the automatic comparison of finger-ridge patterns", *Hughes Laboratory Research Report* 1961, n. 190.

[61] The first fully automated biometric recognition system was the hand geometry based system patented by Robert P. Miller in 1971; reference from Zunkel, R., "Hand geometry based verifications", in A. Jain, et alt. (eds.) *Biometrics: Personal Identification in Networked Society*, 1999.

[62] Fejfar, A./ Myers, J., "The testing of 3 automatic ID verification techniques for entry control", *2nd Int.Conf. on Crime Countermeasures*, Oxford, 25-29 July 1977.

[63] Wildes, R. P., "Iris recognition: an emerging biometric technology", *Proc. IEEE*, 85(9), 1348-1364, 1997; Jain, A./ Bolle, R./ Pankati, S. *Introduction to biometrics*, in Jain, A./ et al. (eds.) *Biometrics: Personal Identification in Networked Society.* Kluwer Academic Press, 1999. Vid. NSTC *Biometrics History*, 2006.

attacks[64] and the approval of the European biometric passport[65]. The latter, at the request of the US, incorporated an automated reading chip that stores biometrics of the face and fingerprints of EU citizens.

In the context of the reflections triggered by this passport, many institutions reflected on the implications of this inclusion. The French Comité Consultatif National d'Éthique pour les Sciences de la vie et de la santé (CCNE), for example, referred in 2007 to the *risk of biometrisation* of the human being[66], while other institutions spoke of hypervigilance, barcodes[67] for humans or biopolitical tattoos[68]. The European Data Protection Supervisor, the Article 29 Working Party, the European Parliament's Committee on Citizens' Freedoms and the experts who for months presented their theses to the UK House of Lords, for their part, also warned of the risks of allowing the use of body-based identification[69]. Under threat of EU citizens being excluded from the visa waiver system for access to the US, however, European legislators had no choice but to accept the imposition of the biometric passport.

2.4. Important limitations of biometric recognition systems: scientific-technical, architectural and social-ethical limitations

In the mid-1990s, a US Treasury report, which was made public some time later[70], highlighted the reasons why the use of biometric technologies had been reserved until practically the 21st century to very exclusive areas of social reality (such as the security of high-level financial operations, or the control of military installations or high national security bodies). One of the milestones that contributed to a change in this paradigm was the Patriot Act, passed in the US in response to the 9/11 attacks. And its consequences prompted a paradigm shift. As many experts foresaw, the inclusion of biometric information in European passports helped to normalise the use of these recognition systems, and they spread uncritically to such everyday contexts as unlocking a mobile phone or accessing a place of entertainment.

This should not, however, obscure the fact that biometric recognition systems, especially those using weak biometrics, still have important limita-

[64] *Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act*, Pub L., No. 107-56, 115 Stat. 272, 2001.

[65] Lyon, D., *La societá sorvegliata*, cit., 2002, 96; Rule, J. B., *Privacy in Peril*, cit., 43-39.

[66] CCNE, *Biométrie, dones identifiants et droits de l'homme*, 2007, cit., 3.

[67] Crews jr., C. W., "Human Bar Code. Monitoring Biometric Technologies in a Free Society", *Policy Analysis*, no. 452, 2002, 1 ff.

[68] Agambe, G., "Bio-political tattooing", *Le Monde*, 11 January 2004.

[69] In detail, Escajedo San-Epifanio, L., *Tecnologías biétricas*, cit., 2017, 71-75, 110 et seq.

[70] Jain, A. K./ Flynn, P./ Ross, A. A.: *Handbook on Biometrics*, Springer, 2008, 1.

tions in their scientific basis, in their IT architecture and have important impacts from a social-ethical and legal perspective.

With regard to their scientific basis, we must bear in mind that, although the human body offers countless possibilities for capturing biometrics, biometric recognition systems are always imperfect[71]. Imperfect, first of all, because there is no universal, permanent or one hundred percent stable biometric[72], nor does it provide an unambiguous basis for a consistent categorisation of human bodies with respect to parameters such as emotions. Secondly, it has to be taken into account that biological variability among people is not homogeneously distributed and that this circumstance makes any recognition system much more efficient with respect to people who are in the average of their identification range than with respect to people who are biometrically atypical. Last but not least, it has to be taken into account that, although many attributes can hypothetically be measured on the human body, the current state of the art does not offer robust measurement methods applicable to all potential sources of information. The human body is not easily biometrisable[73].

As far as computer systems are concerned, it has to be taken into account that the architecture of recognition systems is decisive with regard to parameters such as accuracy, performance and cybersecurity[74]. Even choosing the most optimal algorithms (i.e., algorithms that adequately address the need to capture the uniqueness of the sample)[75], actions such as image enhancement, attribute extraction, matching and decision making are critical elements in all biometric recognition systems. Due to a combination of science-based and system architecture constraints, all biometric recognition systems also have a tolerance threshold that makes a certain range of false positives and false positives foreseeable. Only the human expert eye is able to significantly reduce such thresholds.

[71] De Hert, P./ Scheuers, W./ Brouwer, E., "Machine-readable identity documents with biometric data in the EU -part III- Overview of the legal Framework", *Keesing Journal of Documents and Identity*, 2007/ 22,23-26; Kindt, E., "Biometric applications and the data protection legislation (the legal review and the proportionality test)", *Datenschutz and Datensicherheit (DuD)*, 31/ 2007,166-170; Brouwer, E.R. *Digital borders and real rights: effective remedies for third-country nationals in the Schengen Information System.* Brill, 2008, 137.

[72] Lanitis, A., "A survey of the effects of ageing on biometrical identity verification", *International Journal of Biometrics*, 2 (1), 2010, 34-52.

[73] Magnet, S.A., *When Biometrics Fail. Gender, Race and Technology of Identity,* Duke University Press, 2011, 2.

[74] Maltoni/ Maio/ Jain/ Prabhakar, *Handbook Fingerprint*, 2009, 11-22.

[75] Pfaffenberger, B., *Que's Computer and Internet Dictionary*, Que, 1995, 15; Preneel, B., "An Introduction to Modern Cryptology", in *Cryptology best practices*, KU Leuven, 2018, 19-25.

To this should be added a reflection on the ethical-social and legal impacts of biometric recognition systems. Although at first sight the possibilities of intentional hyper-surveillance seem the most problematic from a fundamental rights perspective, other controversial consequences should not be overlooked, such as: the collateral impact on non-suspects; and the likelihood that biometric information could reveal information about a person's health, ethnic or racial origin or bodily characteristics expressing certain religious convictions (such as the tonsures in the hair, the shaved heads of Buddhist believers or the beards of orthodox Jews, among many others).

Added to this is the fact that many sources of biometric information are exposed and can be collected in a person's daily life with little effort (fingerprints on objects he or she touches, non-consensual images of the face, etc.). Without prejudice to the sensitivity of biometric data relating to sexual identity and sexual orientation, attention should also be drawn to the vulnerable situation of individuals whose body is atypical for a system and of individuals who face the risk of systems collaterally capturing health data associated with their static or dynamic biometric characteristics.

By chance of nature or by events after birth, there are people whose bodies lack the supposedly universal features that support the system or have out-of-range attributes -very large or very small hands, for example- as well as people in whom the information that the system uses as a basis for recognition cannot be adequately captured temporarily -for example, because of a disease. Retinitis, for example, is one of the many eye pathologies that make it difficult to reliably capture the iris pattern. Such circumstances are, *per se*, highly sensitive and stigmatising, and their impact may be exacerbated if people are exposed on a daily basis to biometric recognition systems to which they need to explain that, at least for the time being, their body is not fit to be presented as fresh data. Another group of people at high risk of stigmatisation are those who in screening or categorising applications have, while totally innocent, biometric patterns close to what would have been determined as suspicious profiles.

With regard to health, the literature distinguishes between, on the one hand, direct medical implications and, on the other hand, indirect implications of biometric recognition technologies. Direct implications are those health impacts that can be generated by the use of the components of a biometric recognition system. This is the case of the risk of disease transmission through systems that require physical contact, or the risks of prolonged exposure to capture systems -for example, those that use infrared- especially when the source of biometric information is the iris or retina.

Indirect implications, on the other hand, refer to the possibility that medical information about the individual may be exposed in the operation of the system. Without being exhaustive, some genetic syndromes are expressed in dermatoglyphs, facial features or skin colour. The information most at risk of being collaterally exposed is, in general, that of health conditions that temporarily or permanently prevent the capture of biometric information that serves as the basis for a system. This is the case, for example, of eye conditions that prevent the iris from being seen (macular degeneration, retinitis, retinoblastoma, among others), conditions that have damaged fingerprints and diseases that have substantially altered facial morphology (such as swelling due to mumps, or a dental abscess).

3. Biometric verification, outside the Regulation?

In order to highlight the high risk of real-time remote identification, the White Paper chooses to compare it with the risk it perceives in biometric authentication, which the Commission considers to be much lower risk. This contrast, which is reflected in the White Paper, will be maintained in the different versions of the AIA and has finally prevailed in the final text. This circumstance, however, should not make us forget that for the purposes of the GDPR, both identification systems in the strict sense (remote or otherwise) and authentication systems use biometric data that fall under Article 4.14 and, therefore, are subject to the prohibition set out in Article 9.1[76] regarding the processing of special categories of data. Recently, the CEPD[77] and, following in its wake, the DPAs insist emphatically that the biometric data used in authentication systems (one-to-one) are data that fit the definition of 4.14 of the GDPR, and that they are subject to the special protection and limitations set out in Article 9, because authentication most often aims to identify the person, even if it is "one-to-one". Unless the circumstances and safeguards of Article 9.2 GDPR are met, the processing of biometric identification data is prohibited, even though the AIA may paradoxically consider that some identification systems affected by this prohibition can be certified as suitable to be offered on the market.

It should be noted, therefore, that what will be raised here in relation to verification does not arise from the tensions that have arisen in the develop-

[76] Art. 9.1. RGPD. The processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data intended to uniquely identify a natural person, data concerning health or data concerning the sex life or sexual orientations of a natural person are prohibited.

[77] CEPD, *Guidelines 05/022 on the use of facial recognition technology in the area of Law enforcement*, version 1.0 of 2022 and version 2.0 of 2023.

ment of the AIA with regard to the notion of biometric data in the GDPR, but from the need to determine which forms of biometric recognition would be covered by the AIA regulation and which would not.

An amendment of the European Parliament proposed to incorporate the definition of biometric verification into the AIA, as well as two amendments in recitals 8 and 8a. Given that the notion of remote identification in recital 8 of the proposed regulation included all types of remote identification "*irrespective of technology, processes or types of* data", the Parliament proposes to clarify that "*verification systems that merely compare a person's biometric data with their previously provided biometric data ("one-to-one")"* will be excluded. The reason why verification is excluded from remote identification is striking. Somehow, although it does not explain it, the Parliament seems to presume that the subparagraph on "all kinds of processes" may imply that identification comprises both one-to-many and one-to-one comparisons, but that neither is of concern when they are not performed "remotely".

In 8a, for its part, the Parliament proposes to include a demarcation between remote identification and verification, explaining, in a somewhat confusing sentence, that remote identification systems are distinguished from "close-up" individual verification systems by the purpose for which they are used. Thus, it understands that verification systems only aim to "confirm whether a specific natural person presenting himself for identification" is, for example, authorised to access a service, a device, or premises.

However, the parliament's proposed definition of verification lacks clarity. Note that it says that the person "presents themselves for identification", implying that there is an active participation in their identification, but it does not specify whether as part of this presentation, the person will claim an identity (which would lead to a one-to-one comparison) or if the system should check one-to-many whether the person is actually enrolled or not. For the purposes of data protection law, one-to-one comparison and one-to-many comparison have a very different impact, in particular because one-to-one prevents extensive matching and even allows to operate without the need for a reference database. It will be important, moreover, for Parliament to open the debate on the elements that will delimit verification and remote identification. In this distinction between near and far, the initial position points -without specifying- to a question of physical distance, although the question of the participation of the subject to be identified and whether or not the identifier knows whether or not the subject had previously been enrolled in the system appear in a veiled manner.

Consistent with these recitals, the Parliament proposed to amend point 36 of Article 3 suggested by the Commission by including an indent in brackets to clearly exclude verification from remote biometric identification[78]. Thus, in line with the European Parliament's position, the notion of remote identification system would be described in point 36 as an "*AI system (with the exception of verification systems) intended to identify natural persons remotely by comparing their biometric data with those contained in a reference database, and without the deployer of the AI system knowing in advance whether the person in question will be found in that database and can be identified*".

Both the Council's position and the text finally adopted (Art. 3.41 AIA) go back to the initial concept of biometric identification, without accepting the Parliament's proposed subparagraph. To the same end, and more clearly, what is done is to include in the list in Article 3, point 36, the notion of biometric verification, in these terms "*one-to-one verification, including authentication, of the identity of natural persons by comparing their biometric data to previously provided biometric data* ". Note that automated verification and one-to-one verification appear in a copulative expression, and it is not clear whether they are synonyms, as is the case with the clause "including authentication". What is clear is that, in all cases, biometric data provided in a previous enrolment is indispensable, although it is not stated that these data must be collected "in a reference database". This last clause, on the other hand, will appear in the articles when talking about remote biometric identification.

This notion of verification is particularly relevant since, at various points in the articles, both in the prohibitions in Article 5 and in the high-risk categories in Annex III, it is emphasised that AI systems intended for verification are excluded, depending on the case in question. Thus, for example, for the purposes of the prohibition in Art. 5.1. letter h) it is expressly stated in Recital 17 that, from the concept of remote biometric identification systems "*This excludes AI systems intended to be used for biometric verification, which includes authentication, the sole purpose of which is to confirm that a specific natural person is the person he or she claims to be and to confirm the identity of a natural person for the sole purpose of having access to a service, unlocking a device or having security access to premises* ". Furthermore, point 1 of Annex III, in its final wording -as promoted by the Parliament in its amendments- indicates that verification systems shall not be understood to be included in identification systems for the purposes of Annex III[79].

On the interpretation of these exclusions, if verification is understood to include one-to-one identifications and one-to-many identifications, in the latter case as long as they are not remote identifications, the identification

[78] European Parliament Amendment No 193.

[79] European Parliament Amendments Nos 710 et seq.

technologies excluded from the AIA would be very numerous, which is discussed in the following section.

4. Non-remote identification, in limbo?

The effort to exclude verification by distancing oneself from remote identification contrasts with the lack of attention paid to properly defining non-remote identification (the one-to-many) in the strict sense. The concept of "biometric identification", remote or not, is defined in Recital 15 as "*the automated recognition of physical, physiological and behavioural human features such as the face, eye movement, body shape, voice, prosody, gait, posture, heart rate, blood pressure, odour, keystrokes characteristics, for the purpose of establishing an individual's identity by comparing biometric data of that individual to stored biometric data of individuals in a reference database, irrespective of whether the individual has given its consent or not*". The existence of a reference database, i.e., a many to compare the ones, seems to be the element that differentiates identification from verification. Furthermore, and as already seen in II.3, it seems that the absence of active participation is the differentiating element between "remote" biometric identification and non-remote identification, irrespective of the distance at which it takes place.

The problem arises, as explained in the previous section, from the fact that, apart from the recitals, the articles place non-remote identification as part of verification and therefore under the same status for the purposes of the AIA. The legislators consider that biometric verification is "*likely to have a minor impact on the rights of individuals*", although they do not indicate on what such a likelihood depends. Lumping non-remote biometric identification into the category of verification, in an unclear manner moreover, is a very bad decision from the perspective of upholding citizens' rights and freedoms. Three reasons, moreover, lead us to believe that this was not a well-considered decision.

The decision does not seem very well thought out, firstly, because the AIA, in addition to not explaining the reason for this exclusion when applied to one-to-many identification with the participation of the subject, does not clearly delimit what this *active participation* consists of. This is worrying because the absence of active participation would mean that some cases of non-remote one-to-many identification would fall under the prohibition of Art. 5.1.h or, at least, under the protection of the guarantees provided for high-risk categories. The legislators do not seem to have taken this into account.

Secondly, it does not seem that the legislators have paid much attention to the implications of the use of unique biometric data. It is true that the AIA claims to recall, on more than one occasion, that under the GDPR all unique biometric data belong to the category of sensitive personal data[80]. But it does not seem to act accordingly. The impact of unique identifiers on the rights of individuals depends not only on the alleged active participation of individuals, but also on circumstances such as whether such participation may be against their will, on the particularities of the operational scenario in which they are applied, on the quality and efficiency of the systems or, among others, on the type of decisions or consequences that may result from the use of these systems. The exclusion of these systems from the scope of AIA means, however, that the assessment of their quality and reliability will not enjoy the safeguards that apply to the modalities classified as high risk. Nor does it take into account the fact that one-to-many identification always requires the handling of databases, which is not essential for authentication, since it is possible that persons carry with them (in the passport chip, for example) the stored data that will be used for the comparison with the fresh data.

The biggest inconsistency, however, is with regard to Art. 111 and the large-scale IT systems listed in Annex X of the AIA. Systems like SIS, Eurodac, and others listed in Annex X provide, depending on the situation, one-to-one or one-to-many identification utilities, currently with the active participation of the subjects. None of them function, to date, as a remote identification system, unless remote identification can be understood as searches using fingerprints not taken from the body (but present on physical objects in which they are latent) or, as the case may be, images of the person that have been taken for purposes other than enrolment in the strict sense. In most cases, especially at border crossings, fingerprints and faces of persons are captured with their active participation, in the sense that the capture of quality fresh biometrics requires that the person is exposed to the system in very specific ways. Thus, for example, in many cases it is essential that people are willing to roll their fingers or press the scanner to obtain rolled or flat prints, or to place their face at a certain angle to capture the geometry of the face, having removed lenses or strands of hair that could partially conceal it. Only in databases for police use are searches sometimes carried out from images or fingerprint captures taken outside the enrolment scenario, although we must bear in mind that Art. 2.3 excludes from its application biometric recognition that Member States carry out for military, defence or national security purposes, although the GDPR is very strict in the guarantees it requires

[80] This is reiterated in recital 54 of the final text, although it goes on to state that the fact that they constitute a category of sensitive data leads to classify ‘several’, but not all, cases of ‘critical use’ of remote biometric identification systems as high risk.

of them. The interpretation of the whole is therefore far from clear. It is true, in any case, that if a systematic interpretation excludes de facto the possibility of applying the AIA to Annex X biometric systems, one might wonder what has motivated its inclusion.

5. Is biometric screening covered?

Biometric screening systems are systems that recognise "something", some characteristic -identifying or non-identifying- in a set of people captured in the moment, without any prior process of individual enrolment.

These can be categorising systems, for example, that allow to calculate how many people of each age and sex range are present in a space, as well as systems that aim to locate in a crowd either suspicious behaviour or possible similarities between the people present and artificial biometric patterns, created in a robot-like manner as being close to those of persons wanted for committing criminal acts. Close to, it should be noted, this implies that such artificial patterns have under no circumstances been created by enrolment.

As they do not operate on enrolled identities, it is clear that they cannot be understood as falling under the concepts of identification (remote or not) or authentication or verification. It is possible, however, to understand that screening, regardless of the reference of the screen, can fit into the idea of biometric categorisation. On that basis, and defending a guarantee-based interpretation of the AIA, it seems possible to argue that, depending on the categories used, AI screening systems will, in most cases, fall under the prohibitions of Article 5 or, where appropriate, under the high-risk modalities of Annex III.

III. Biometric recognition systems affected by the prohibitions and restrictions of Article 5 of the Regulation: social assessment, prediction of criminal dangerousness, extension of facial recognition databases, inference of emotions in certain contexts and categorisations on specially protected data under Article 9 RGPD.

1. Systematics of Article 5 regarding biometric recognition

The proposals of the Commission, the Parliament and the Council distributed differently the forms of biometric recognition that, respectively, should be prohibited or, as the case may be, be included in the category of high-risk forms. The final version of the text yields in some respects to the demands of the Parliament, but the concessions made to the interests of the Member States are also relevant.

1.1. The Commission's initial proposal

Article 5 of the Commission proposal included in its paragraph 1 the prohibition of four AI practices, relating to the use of certain subliminal techniques (letter a), the substantial alteration of the behaviour of vulnerable persons (letter b), the assessment of the reliability of natural persons, based on their known or predicted social behaviour or personal characteristics, and (letter d) the use of real-time remote biometric identification systems in publicly accessible areas for law enforcement purposes. Given that some of the cases mentioned are not directly related to biometric recognition (especially letters a and b) and have been dealt with by other authors of this Treaty, the scope of all these proposed prohibitions will not be analysed here, especially since the Commission's proposal is far from the text finally adopted.

The aim of this section of the commentary is to situate the debate during the period when the text was being drafted, and then to go into the details of the regulation that was finally approved.

1.2 Parliament's stance

The Commission's proposal was criticised[81] for disregarding relevant previous positions, including those of the Council of Europe (2021)[82], the EU Agency for Fundamental Rights (2019)[83] and the Parliament[84]. In these documents it had been expressed very forcefully that biometric surveillance, emotion recognition or categorisation have a very negative impact on a wide range of fundamental rights, as well as on the principles of the rule of law and democratic values[85]. Further criticism was added to the proposal due to

[81] Barkane, I., "Questioning the EU Proposal for an Artificial Intelligence Act: The Need for Prohibitions and a Stricter Approach to Biometric Surveillance", *Information Polity* 27, 2022: 147-162.

[82] Council of Europe, *Guidelines on Facial Recognition*, 2021.

[83] FRA, *Facial recognition technology: fundamental rights considerations in the context of law enforcement*, 2019.

[84] Madiega, T./ Mildebrath, H., *Regulating facial recognition in the EU*, European Parliament, 2021.

[85] See critiques, among others, by Veale, M/ Zuiderveen Borgesius, F., "Demystifying the Draft EU Artificial Intelligence Act", *Computer Law Review International*, 2021, 22: (4), 97-112; Malgieri, G/ Ienca, M., "The EU regulates AI but forgets to protect our mind", *European Law Blog*; EDRI, *New ECI calls Europeans to stand together for a future free from harmful biometric mass surveillance*, 2021; EDPB, EDPS, *Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act), 2021.*

the fact that the alleged human rights approach applied to the proposal had not been translated into effective mechanisms to address the most threatening assumptions[86].

Along these lines, and in line with previously mentioned Resolutions, the European Parliament was in favour of prohibiting or, as the case may be, restricting the use of biometric recognition technologies as much as possible. Beforehand, the operational scenario in which these technologies were intended to be used or the purpose of the recognition did not overly qualify the Parliament's rejection, showing in its position very little intention to accept exceptions to the general restrictions. Thus, its proposed amendments include new prohibitions that clearly imply the use of biometric recognition modalities and propose that the Commission's proposed point (d) be worded more broadly. In particular, in its amendment 220 the European Parliament proposes to prohibit[87] *"d) The use of "real-time" remote biometric identification systems in publicly accessible areas*" with very few exceptions.

Moreover, in its amendment 224 and the following ones, the Parliament also proposed that certain cases described by the Commission as high risk should be relocated among the prohibited AI practices. This is the case, for example, of assessments based on personality that aim to predict criminal dangerousness or the risk of recidivism, of certain activities related to facial recognition database extension systems, or of emotion recognition in workplaces, educational establishments and at borders.

1.3. The Council's position and the finally adopted text

Some Member States, in particular Germany, France and Italy, were against the AIA being able to limit in any way the use of remote biometric identification systems for national security purposes. Certainly, the GDPR severely limited the possibility to use such systems, but with the support of national parliaments, especially in the form of Law, it was possible to invoke the exceptions of Art. 9.2. RGPD.

This vision of the Member States, however, was far removed from the previous position of the Parliament, which was endorsed in its position on the AIA. In order to prevent this disagreement from becoming an insurmountable obstacle in the AIA process, the Council opted for an initial po-

[86] Smuha, N.A., "Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea", *Philosophy & Technology*, 2021, 34: 91-104; Mantelero, A./ Esposito, M. S. "An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems" *Computer Law & Security Review*, 2021: 41.

[87] Amendments adopted by the European Parliament on 14 June 2023, cit. supra.

sition aligned with the Commission[88], leaving for the trilogue period its push to minimise those prohibitions that could affect the defence and security interests of Member States. The broadening of the verification notion in the AIA and the surprising new paragraph of Art. 2 (Art. 2.3 AIA)[89] marked an important step in this direction, together with a proposal for a new wording of the exceptions applicable to prohibited AI practices in Art. 5.

This set of prohibitions, in particular the one concerning certain uses of remote biometric identification systems, has often been pointed to as a symbol of the supposedly guaranteeing nature of the AIA, although the reality is far from this assertion[90]. In addition to the ambiguity of some of the exclusions, which are many, the technical quality of some of the approved rules does not make it possible to foresee a significant instrumental effectiveness with respect to the legal system already in force. It is therefore striking that the reference to prohibited practices continues to be a persistent topic when communicating to citizens that European legislators have reflected in the AIA a strong commitment to justice, security and other relevant values[91].

Within this set of "*practices to be prohibited*" -but not exactly prohibited- the interest raised by remote biometric recognition in real time and in publicly accessible spaces, carried out with the aim of ensuring the application of the law, has been particularly noteworthy. The exceptions and the need for clarification of the prohibition were notable -and divergent- in the positions taken by the Commission, the Parliament and the Council. The definitions in Article 3 and a large part of Article 5, including paragraphs 2 to 8, also occupy a considerable place in the final text, spread over the recitals. For this reason, it has been considered appropriate to break down the treatment of biometric recognition systems affected by the prohibitions in Chapter II into two separate headings. This section deals with biometric recognition systems covered by Article 5.1 (c) to (g). AIA, leaving for the next section the singular treatment of remote biometric identification covered by Art. 5.1(h), together with paragraphs 5.2. to 5.8.

[88] See *the Fourth Presidency's compromise text*, 19 October 2022, cit. supra.

[89] See above the full text, at I.1.

[90] Cotino Hueso, L., "Sistemas de inteligencia artificial con reconocimiento facial y datos biétricos, Mejor regular bien que prohibir mal", *El Cronista del Estado Social y Democrático de Derecho,* n.º 100, 68-79, 73.

[91] García-Villegas M., "The Symbolic Uses of Law: At the Heart of a Political Sociology of Law", in *The Powers of Law: A Comparative Analysis of Sociopolitical Legal Studies*. Cambridge University Press, 2018, 19-37.

2. Biometrics affected by the prohibitions of Art. 5.1 (C, D and E) of the Regulation: personality assessments for the purposes of citizen scoring, crime risk prediction and extension of facial recognition bases.

Letters (c), (d) and (e) of the final AIA text contain factual scenarios in which biometric technologies may fit, although these do not exhaust the full range of possible scenarios. Point (c) would cover biometric technologies that can be used to assess inferred or predicted behaviour or personality characteristics. Point (d) refers to AI systems, including biometric-based systems, that can be used to perform risk assessments of natural persons in order to assess or predict the risk of a natural person committing a crime, provided that the form of the reference profiles is an assessment of personality traits and characteristics. Some soft biometrics are used, not without controversy, for such purposes and merit considerations from the perspective of the principles of the democratic rule of law and criminal policy, which are dealt with in more detail in another chapter of this Treaty[92]. The risk of a resurgence of *biologisations of criminality,* such as the mythical thesis of the *uomo delinquentis* of Lombroso and his disciples[93], disguised or hidden under the haze of the apparent infallibility of digital technologies, is a circumstance that should not be overlooked.

A caveat is provided for the case where the assessment is not merely predictive, but is made -supported, if necessary, by AI- on the basis of a person's involvement in criminal activity and "on *the basis of objective and verifiable facts directly related to criminal activity*".

In the case of letter e), on the other hand, it is clear that the creation or extension of facial recognition databases requires -either by enrolling persons or by non-consensual capturing of facial images- the use of biometric technologies. The prohibition, however, extends to two very different cases, given that in one case it refers to images that are supposedly broadcasted openly, with a very different legal regime depending on whether or not the owner has consented to such dissemination, and in the second case, it refers to closed circuits, pointing to private recordings (not surveillance). It should also be noted that in both cases what is prohibited is the "non-selective" extraction of images, a concept that is not specified but which clearly leaves selective extraction outside the prohibition.

[92] See, in this work, the work of F. Miró Llinares on predictive policing systems and emotion recognition systems.

[93] Wechsler, H., „Biometric Security and Privacy Using Smart Identity", *Review of Policy Research*, vol. 29, 1/ 2012, 78-79; Strasser, P., „Biometrie – ein Schritt in die Überwachungsdemokratie?", Schaar, P. (ed), *Biometrie Und Datenschutz – Die vermessene Mensch*, 2007, 14-15.

3. Biometric emotion recognition in workplaces and educational establishments, except for medical or security purposes (art. 5.1. f). Concept of emotion recognition in the Regulation

For centuries, law enforcement agencies have used faces not only to identify but also to try to read mental states and infer suspicious behaviour[94]. Despite the impression that may be given by the recent attention that face recognition is receiving in the political space, academic literature and even the press, it is a long-standing area of technological development -at least since the 1970s[95],[96]. It is true, however, that advances in machine learning and improved computer vision techniques, combined with biometrics, have significantly increased the surveillance capabilities of police forces[97], with the risk of perpetuating inappropriate forms of profiling and reinforcing categories of suspicion on certain groups of subjects[98].

The EDPS, prior to the final adoption of the AIA, called for stricter regulation of facial recognition technologies and the use of biometric recognition systems in a broad sense[99], including '*gait, fingerprints, DNA, voice, keystrokes and other biometric or behavioural signals*' when used by law enforcement, whether in public or other spaces[100]. Attention to the use of such systems by businesses

[94] Miranda, D., "Identifying Suspicious Bodies? Historically Tracing the Trajectory of Criminal Identification Technologies in Portugal", Surveillance & Society 2020, 18 (1): 30-47.

[95] Gray, M. "Urban Surveillance and Panopticism: Will We Recognize the Facial Recognition Society?", *Surveillance & Society*, 2003, 1(3), 314-30; Introna, L., Wood, D. "Picturing Algorithmic Surveillance: The Politics of Facial Recognition Systems" *Surveillance & Society* 2004 (2), 177-98.

[96] Urquhart, L./ Miranda, D., "Policing faces: the present and future of intelligent facial surveillance", *Information & Communications Technology Law*, 2021, 31(2), 194-219.

[97] Kotsoglou, K. / Oswald, M. "The Long Arm of the Algorithm? Automated Facial Recognition as Evidence and Trigger for Police Intervention" (2020) 2 Forensic Science International: Synergy 86-89; Venema, R. "How to Govern Visibility? Legitimizations and Contestations of Visual Data Practices after the 2017 G20 Summit in Hamburg", *Surveillance & Society* 2020, 18 (4) 522-39; Purshouse, J. / Campbell, L., "Privacy, Crime Control and Police Use of Automated Facial Recognition Technology", *Criminal Law Review* 2019 (3), 188-204.

[98] Garvie, C./ Bedoya, A. /Frankle, J., *The Perpetual Line-up: Unregulated Police Face Recognition in America*, Georgetown Law, Center on Privacy & Technology, 2016; Williams, D., "Fitting the Description: Historical and Sociotechnical Elements of Facial Recognition and Anti-Black Surveillance", *Journal of Responsible Innovation*, 2020, 7 (1), 74-83.

[99] European Data Protection Board and European Data Protection Supervisor, (EDPB-EDPS), *Joint Opinion 5/2021 on the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act*), 2021.

[100] European Parliament, *Motion for a European Parliament Resolution on Artificial Intelligence in Criminal Law and its Use by the Police and Judicial Authorities in Criminal Matters*, 2021.

has gone more unnoticed, although they are increasingly used in advertising and commercial contexts[101]. Behavioural data is often aggregated with preference or purchase data[102], as well as data that can assist in crime prevention[103]. Such uses are also being extended to workspaces, for the time being in a rather unlawful manner. Unlike video surveillance, digital monitoring results in often unnoticeable but pervasive surveillance[104] and we actually know very little about how *big data* is used in this kind of surveillance[105].

Emotion recognition is an interdisciplinary research field that encompasses, among others, knowledge from the disciplines of psychology, cognitive science and computer science.[106]. It aims at enabling computers to capture human emotions and intentions and manage them for different purposes[107]. Some automated facial recognition systems, for example, are trained to detect six basic emotions (anger, joy, fear, surprise, disgust and sadness[108]) either in static form (forehead wrinkles, chin or eyebrow position)[109], or by capturing biometric information from different micro-expressions over a period of time (blink rate, chin or eyebrow movements, gaze fixation, etc.).

According to the Commission proposal, an emotion recognition system is '*an AI system that aims to identify or infer emotions or intentions of natural persons from their biometric data*'. This type of system may use biometric data that is unique and falls under Art. 4.14 GDPR, but it usually uses non-identifying data, as explained in II.1 above. Thus, an 'emotion recognition system' is defined as an '*AI system designed to distinguish or infer the emotions or intentions of natural persons from their biometric data* (Art. 3.39 AIA)', the latter being understood as the biometric data described in No. 34 of the same article.

It should be noted that the prohibition only covers the placing on the

[101] McStay, A. *Emotional AI,* Sage, 2018; Stark, L., / Huey, J. "The Ethics of Emotion in AI Systems", *FAccT 21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, 782-93.

[102] Lyon, D., *Surveillance after Snowden*, John Wiley & Sons, 2015, 76-77.

[103] Brayne, S., "Big data surveillance: The case of policing". *American Sociological Review* 82(5), 2017, 977.

[104] Van Oort, M., "The Emotional Labor of Surveillance: Digital Control in Fast Fashion Retail", *Critical Sociology*, 2019, 45(7-8), 1167-1179.

[105] Brayne, S., "Big data surveillance", cit. 2017, 977.

[106] De Gregorio, G., "The Rise of Digital Constitutionalism in the European Union", *International Journal of Constitutional Law* 19.1, 41-70, 2021.

[107] Picard, R. W., "Affective Computing: Challenges", *International Journal of Human-computer Studies* 59.1, 2003, 55-64.

[108] Lewinski, P./ Trzaskowski, J./ Luzak, J., "Face and Emotion Recognition on Commercial Property under EU Data Protection Law", *Psychology & Marketing* 33 (9), 2016, 729-46.

[109] Eckman, P., *Emotions Revealed: Understanding Faces and Feelings*, HB, 2003, p.17-19.

market, putting into service for this specific purpose, or the use of *AI systems to infer the emotions of a natural person in workplaces and educational establishments*, except where the AI system is intended to be installed or placed on the market for medical or security reasons. It was taken out of the final text from the Parliament's proposal that these systems could be used for border control, which could be affected by Articles 111 et seq. or the exclusion of Art. 2.3 AIA. Other cases, not covered by the prohibition or excluded in the above-mentioned articles, are in principle included in the set of high-risk AI modalities in Annex III.

Attention should be drawn, however, to the fact that, although the legislators have acknowledged their concern about the significant margin of imprecision in the recognition of emotions (and intentions), no moratorium on the possibility of introducing this type of system on the market has been envisaged. Thus, Recital 44 refers to "serious concerns" about the scientific basis of such systems, pointing to *limited reliability, lack of specificity and limited possibility of generalising* as the main shortcomings. In contexts such as employment or education, which are covered by the ban, this low reliability exacerbates situations of imbalance of power characteristic of these areas, but in general, legislators should have taken into account that, due to their little or no potential to identify unique individuals, the collection of these biometric data is excluded from Article 4.14 GDPR (and the guarantees of Article 9) and needed to obtain additional guarantees from the AIA.

4. Biometric categorisation in order to deduce or infer their race, political opinions, trade union membership, religious or philosophical convictions, sexual life or sexual orientation (art. 5.1 G).

Article 5.1 (g) prohibits the placing on the market, putting into service for this specific purpose or the use of *biometric categorisation systems that individually classify natural persons on the basis of their biometric data in order to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation.* This prohibition, it states in its final paragraph, does not include the tagging or filtering of lawfully acquired biometric data sets, such as images, or the categorisation of biometric data in the field of law enforcement. With regard to the first exclusion, the labelling of lawfully acquired data, a number of doubts arise with regard to its interpretation. Biometric categorisation is not based in principle on individualising data, which fall within the notion of Art. 4.14 GDPR, so it is to be expected that the exclusion to the prohibition will remain within the modalities of categorisation. The reference to labelling, however, mentions previously lawfully acquired data and includes images, so it is not entirely clear that the applicability of the guarantees of the GDPR cannot be ruled out.

It is not clear, to be honest, under what circumstances an image could have been lawfully obtained to label a person in order to infer their political opinion, trade union membership, or religious beliefs; perhaps a photo published by a political association on the occasion of a celebration, or images obtained at the celebrations of public demonstrations of various kinds? Frankly, the legal basis for storing and labelling such information is doubtful, as is the possibility of having a legal basis, and of justifying in a democratic State the necessity and usefulness of processing such sensitive data. In addition, of course, there is little chance that biometric profiling -which is what the article is about- will actually capture some of this information.

Thus, and in accordance with the current state of the art, the reference to the possibility of inferring political opinions, trade union membership or philosophical or religious beliefs from biometric data of individuals by means of biometric categories is striking, and the reliability of the systems that have been developed with regard to sexual orientation and sexual life is somewhat dubious. Skin colour or certain attributes clearly present on the body, such as hair type or certain striking facial features, also do not allow us to categorise with the quality that some voices seem to presuppose, because miscegenation in contemporary societies has grown considerably. Strikingly, on the other hand, from the list of special category data in Art. 9.1 GDPR has excluded from the prohibition biometric categorisations that seek to deduce or infer the health conditions of individuals. To this day, largely because they are used as a reference in the healthcare field, these are the categories where the most scientific-technological effort is being made.

As a final note, it should be remarked that the text finally adopted, contained in Art. 3.35 AIA, defines a "biometric categorisation system*"* as *"an AI system designed to place natural persons in specific categories on the basis of their biometric data",* although it is to be criticised that systems of this type which are accessory to a commercial service, and strictly necessary for objective technical reasons, have been excluded from such a definition. It could be, for example, a system that records the age of persons for the purpose of controlling the exclusion of sales of tobacco or alcohol to minors, but this does not explain why they are excluded from the definition of biometric categorisation system. They should have been excluded, as appropriate, from the literal wording of the prohibition or the status, if any, of high-risk categories.

IV. Biometric recognition systems affected by the prohibitions and restrictions of Article 5 Regulation (and II): "real-time" remote biometric identification in public access areas for law enforcement purposes

1. Global concern about remote biometric identification in public access spaces

The use of facial recognition technology is spreading across the globe[110]. Africa, Latin America and Asia (especially China)[111] are often cited, but EU Member States are also resorting to remote biometric recognition, even if only occasionally in emergencies or in sensitive scenarios of mass events[112]. Globally, civil society organisations are increasingly calling for regulation of the conditions under which AI-driven biometric identification, especially remote biometric identification, can be deployed[113]. In places with inconsistent (or non-existent) rule of law, the potential impact of such police uses on the rights of individuals is greater[114], but this means that the uses by corporate and business forces, exploiting loopholes in the law, sometimes go unnoticed[115].

A high percentage of this type of identification uses facial biometrics[116] and, with them, they aim to identify -or at least locate- suspicious persons.

[110] Commission Nationale de l'Informatique et des Libertés (CNIL), *Reconnaissance Faciale – Pour Un Debat À la Hauteur des Enjeux*, 2019, 3; Urquhart, L./ Miranda, D., «Policing faces: the present and future of intelligent facial surveillance», Information & Communications Technology Law, 31(2), 2021, 194-219.

[111] Dauvergne, P. "Facial recognition technology for policing and surveillance in the Global South: a call for bans", *Third World Quarterly*, 43(9), 2022, 2325-2335.

[112] European Data Protection Board, *Guidelines 05/2022 on the Use of Facial Recognition Technology in the Area of Law Enforcement,* 2022, 7 et seq.

[113] Ala-Pietilä, P./ Smuha, Nathalie A., "A Framework for Global Cooperation on Artificial Intelligence and its Governance", in *Reflections on Artificial Intelligence for Humanity*, B. Braunschweig/ M. Ghallab (eds.), Springer, 2021, 253,254.

[114] Zalnieriute, M., "Facial recognition surveillance and public space: protecting protest movements", *International Review of Law, Computers & Technology*, 2024, 1-20; O'Flaherty, M., "Opinions, Facial Recognition Technology and Fundamental Rights", *European Data Protection Law Review*, 2020, 6 (2), 170 et seq.

[115] Dushi, D., "The Use of Facial Recognition Technology in EU Law Enforcement: Fundamental Rights Implications", *Global Campus South East Europe*, 2020, 4; Raposo, V. L., "(Do Not) Remember My Face: Uses of Facial Recognition Technology in Light of the General Data Protection Regulation", *Information & Communications Technology Law* 45, 2022, 32 (1).

[116] Negri, P./ Hupont, I./ Gomez, E., "A Framework for Assessing Proportionate Intervention with Face Recognition Systems in Real-Life Scenarios", *Computers and Society*, 2024 (2), 12.

However, during the COVID-19 pandemic, they were also used for public health purposes[117], with notable uses in countries such as Russia and China[118]. The legal situation of their use in Europe, even after the approval of the AIA, is, to say the least, ambiguous[119], in view of the enormous patchwork of primary and secondary laws of the EU or Member States that regulate some aspects of this type of recognition, in addition to the resolutions and guidelines of different institutions.

The Italian Guarantor ruled in 2021 on the SARI Real-time mobile system, designed to be installed in a specific location and to analyse in real time the faces -maximum capacity 10,000- filmed in a delimited geographical area and equipped with a series of interconnected cameras[120]. If, through a facial recognition algorithm, SARI finds a match between a face present on the watch list and a face filmed by one of the cameras, the system is able to generate an alert that attracts the attention of operators, although in the meantime it is able to record video streams -like traditional video surveillance systems-.

In its resolution, the Guarantor is not only concerned about the persons included on watch lists, but also about the collateral surveillance that this type of system generates with respect to persons present in public spaces or participating in political or social demonstrations who, beforehand, have not been listed by the police forces as persons who are the object of attention. The Guarantor considers that the legal basis for applying a system of these characteristics does not exist, after analysing both the GDPR and precepts of the Italian Code of Criminal Procedure, among others.

In Spain, to cite another example, a pilot system installed in MERCADONA -without prior consultation with the AEPD or a data protection impact assessment, which is mandatory under art. 35.1 of the GDPR- received a sanction for violation of the lawfulness of the processing, aggravated by the fact that special categories of data were involved[121].

[117] Raposo, V. L., / Du, L., "Facial recognition technology: is it ready to be used in public health surveillance?", *International Data Privacy Law*, 14 (1), 2024, 66-86; Raposo, V. L., "Can China's "Standard of Care" for COVID-19 Be Replicated in Europe?", *Journal of Medical Ethics* 46, 2020, 451.

[118] Article 19, "Emotional Entanglement: China's Emotion Recognition Market and its Implications for Human Rights", 2021.

[119] Raposo, V. L., "Look at the camera and say cheese: the existing European legal framework for facial recognition technology in criminal investigations", *Information & Communications Technology Law*, 33(1), 2023, 1-20.

[120] Garante per la Protezione dei Dati Personali, *Parere sul Sistema Sari Real Time*, doc. 9575877, n.º 127 of 25 March 2021.

[121] A detailed analysis of this resolution can be found in Simón Castellano, P. / Dorado

The prohibition contained in Article 5 AIA and the remote identification modalities classified as high risk do not alleviate this situation of lack of clarity, given that, as mentioned above, Article 2.3 excludes from the scope of application of AIA both systems for military use and those used in the context of national security, as well as those whose sole purpose is scientific research and development.

2. Interpretation of the key concepts of the prohibition in Art. 5.1.h)

2.1. Real-time and delayed remote biometric identification

As mentioned above, the notion of remote biometric identification was not discussed at length. The intention was to adopt a functional concept -*see* Recital 8- so that the type of technology and the specific processes or types of data used for this purpose are not central to the definition. What is relevant is that the active participation of persons is not required (Art. 3.34). However, the reference in Recital 8 to the fact that the comparison with the reference database is carried out "*without knowing in advance whether the person in question will be in that database and can be identified*" is questionable, because in the event that the identifier "knows" in advance -it is not known in what terms- that the person will be in that database, such a clause would lead to considering that the identification is not remote. Insofar as the text does not include this provision, and given the legal value of the recitals -in comparison with the text of the articles- it is convenient for remote identification to be delimited by sticking exclusively to Art. 3.34 AIA.

Much more controversial was the question of distinguishing between systems that identify in real time and systems that identify in delayed mode. The recitals indicate that this nuance translates into differences both in terms of characteristics and in the forms of use and risks involved, which is why an effort is made to proceed to delimit each of these concepts. The final text indicates that real-time identification systems are those in which the three phases of this type of model, i.e., collection of biometric data, comparison and identification, *"occur instantaneously, almost instantaneously or, in any case, without significant delay"*. In case of doubt, any remote biometric identification system that cannot be considered a real-time system will, by default, be considered a delayed system for the purposes of the AIA and therefore subject to a much more flexible regime in the joint application of points 42 and 43, which cover real-time and delayed remote identification systems respectively.

Ferrer, X., "Limites y garantías constitucionales frente a la identificación biétrica", *Revista de Internet, Derecho y Política – IDP*, n.º 35, March 2022, 1-13.

With respect to traditional video surveillance, we would therefore be dealing with systems that, in real time, are capturing under their surveillance range and, at the same time, matching biometric patterns captured in real time with patterns stored in an information base. On the other hand, in "delayed" systems, the fresh data would be data already collected, and the matching would have to take place with a significant delay, at least to the point where it could not be considered as acting "almost live". This delimitation does not really provide much of a guarantee. It would have been better, for example, to require that fresh capture and matching be somehow balkanised, with no possibility for the system to start the matching process autonomously and without minimal human intervention.

2.2. Operational scenario and mission: public access spaces and law enforcement purposes

The prohibition of Article 5.1 (h) applies only in cases where the identification system, in addition to being remote and operating in real time, is located in a publicly accessible area. The final text of the AIA describes a "publicly accessible space*"* as "*any publicly or privately owned physical place accessible to an undetermined number of natural persons" (point 44, Art. 3)* and does not consider it relevant whether or not certain conditions for access -such as having a ticket in the case of a museum or a theatre- must be met, or whether the area has possible capacity or security restrictions, including age restrictions. There is no doubt, however, that the space must be "physical", so that online spaces, regardless of how they can be accessed, are not the object of attention for the purposes of this precept.

The proposed recitals of the Commission, Parliament and Council have considerably increased the list of sites that may meet these characteristics[122], and some remarks have been made to clarify cases that remain in doubt. It is recognised, however, that whether or not a site is publicly accessible will have to be determined on a case-by-case basis, taking into account the particularities of each specific situation. As a guideline, it is explained that the fact that access is physically possible (because a gate is open) does not imply, by itself, that the space is publicly accessible, as there may be indications -such as a sign- that access is restricted. Nor are business and factory premises or places to which only employees or service providers are intended to have access, or public access areas in prisons, public access. It is also noted that some spaces

[122] Without being exhaustive, the list includes e.g., shops, restaurants, cafés; services, e.g., banks, professional activities, catering; sports, e.g., swimming pools, gyms, stadiums; transport, e.g., bus, underground and railway stations, airports, means of transport; entertainment, e.g., cinemas, theatres, museums, concert halls, conference halls; leisure or other, e.g., public roads and squares, parks, forests, children's playgrounds. See recital 19 of the adopted text.

may contain, on the one hand, public access areas -such as lobbies- and, on the other hand, non-public access areas.

The term "for law enforcement purposes", on the other hand, was incorporated in the final drafting process to replace the previous reference to the application of the Law. Previously, the Commission, Parliament and Council had maintained the reference to "law enforcement purposes" in their positions, but this term probably became meaningless when Art. 2.3 AIA excluded the use of AI systems by Member States for military, defence and national security purposes from the scope of application.

The term law enforcement is further developed in Recital 24, in addition to two definitions in Article 3. One definition is found in Art. 3.46 AIA, according to which law enforcement means activities carried out by or on behalf of law enforcement authorities "*or the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including safeguarding against and preventing threats to public security*". The second definition is contained in Art. 3.45 AIA, which defines law enforcement authorities as those public authorities competent for the activities just listed as part of Art. 3.46 AIA, including any body or entity entrusted by the law of a Member State with the exercise of the authority and powers necessary to carry out such activities.

2.3. Exceptional to the extent that the use is strictly necessary to achieve one or more of the following objectives

The inclusion of some forms of remote biometric identification among the prohibitions in Article 5 seems to be more politically motivated than a matter of legislative technique, given that the exceptions to this prohibition have been so important since the Commission's proposal. The way in which the text, beset by so many exceptions, was articulated was very unclear. It is number 21 of the recitals that most honestly expresses what the actual provision is -not prohibition- with regard to these systems: "*Each use of a 'real-time' remote biometric identification system in publicly accessible spaces for the purpose of law enforcement should be subject to an express and specific authorisation by a judicial authority or by an independent administrative authority of a Member State whose decision is binding* ". In other words, it is the unauthorised uses to which the AIA will pay attention, although it is questionable whether or not a regulation aimed at assessing the risks of allowing AI systems to be placed on the market is the right place to refer to these cases. As a matter of common sense, in a state governed by the rule of law, surveillance of this kind without a legal basis and without proper authorisation is contrary to the constitutional systems of the Member States, and it is unnecessary to devote so much space to it in the AIA.

The recital goes on to state that "such authorisation should, in principle, be obtained prior to the use of the AI system with a view to identifying a person or persons. Exceptions to that rule should be allowed in duly justified situations on grounds of urgency, namely in situations where the need to use the systems concerned is such as to make it effectively and objectively impossible to obtain an authorisation before commencing the use of the AI system. In such situations of urgency, the use of the AI system should be restricted to the absolute minimum necessary and should be subject to appropriate safeguards and conditions, as determined in national law and specified in the context of each individual urgent use case by the law enforcement authority itself." This is little new with respect to what has already been said about the constitutional provisions. It should be noted that we are not talking about the marketing of the systems but about their use, so that going into specifying in a Regulation such as this -as if they were exceptions to a prohibition that is not such- in what circumstances Member States may or may not grant such authorisations under their domestic law for the purpose of applying the Law was foreseen, as it has been, a fertile ground for interminable discussions.

Furthermore, recital 22 recalls that "*within the exhaustive framework set by this Regulation that such use in the territory of a Member State in accordance with this Regulation should only be possible where and in as far as the Member State concerned has decided to expressly provide for the possibility to authorise such use in its detailed rules of national law. Consequently, Member States remain free under this Regulation not to provide for such a possibility at all or to only provide for such a possibility in respect of some of the objectives capable of justifying authorised use identified in this Regulation*".

All these additional explanations have been necessary because the text finally adopted moves away from the European Parliament's amendment proposal for a comprehensive ban on biometric identification systems, at least in its application by public authorities. Such a text would have been simpler to draft, although questionable also in terms of the necessity and timeliness of its inclusion in the AIA. Parliament, in its amendment 330, proposed to prohibit all use of remote biometric identification systems -then in Article 5.1 (d) – removing -and making unnecessary the interpretation of- the three situations currently covered by Article 5.1 (h).

Parliament's amendment was not successful and the current text allows for exemptions from the prohibition in point (h) for the use of real-time remote biometric systems in public space for the pursuit of certain objectives:

(a) the search for possible victims of crime, including missing children;

(b) responding to specific threats to the life or physical safety of natural persons, or threats of terrorist attacks;

and (c) the localisation or identification of suspects of the offences listed in Annex II[123], coinciding with those of Council Framework Decision 2002/584/JHA, provided that the penalty foreseen for such offences in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least four years.

3. Paragraphs 2 to 8

Paragraphs 2 to 8 of Article 5 are exclusively concerned with completing the exceptions to the prohibition contained in Article 5.1 (h). As a political option, it was seen as preferable to symbolically include the prohibition of remote biometric recognition in the list in Article 5.1, even if a huge effort was then required to specify the minimum -or null- scope for the actual application of this prohibition.

Before going into detail, it is worth making an observation. National security and defence are the most favourable areas for the use of remote identification biometrics, but Art. 2.3 potentially excludes national ones and Art. 111 AIA opens the way to exempt those related to border control. We have also said that some systems used in national security and border control are non-remote identification systems (one-to-many) and, given the way the term verification is defined, it seems that the same legal status is envisaged for them as for one-to-one identification.

In view of this, was there really any need for such a disorderly drafting of point (h), which begins rather bluntly only to be progressively emptied of content in its subparagraphs, and then in Article 5.2 to 8? Moreover, it should be added that some provisions are nothing more than reiterations of provisions already laid down in other legislation. In this regard, paragraph 8 (Art. 5.8 AIA) recalls that "*this Article shall not affect the prohibitions that apply where an AI practice infringes other Union law* ".

Article 5.2 indicates the need to take into account a number of aspects when deploying a remote biometric identification system for the purposes that, exceptionally, Article 5.1 (h) has said it is permitted to do so. The aspects are, firstly, a) the nature of the situation giving rise to the possible use, and in

[123] Annex II: *List of offences referred to in Article 5(1), first subparagraph, point (h)(iii):* terrorism; trafficking in human beings; sexual exploitation of children and child pornography; illicit trafficking in narcotic drugs or psychotropic substances; illicit trafficking in arms, munitions and explosives; murder, assault and battery; illicit trafficking in human organs or tissues; illicit trafficking in nuclear or radioactive materials; kidnapping, illegal detention or hostage-taking; offences within the jurisdiction of the International Criminal Court; hijacking of aircraft or ships; rape; environmental crime; organised or armed robbery; sabotage; participation in a criminal organisation involved in one or more of the offences listed above.

particular the seriousness, likelihood and magnitude of the harm that would occur if the system were not used; and, secondly, b) the consequences that the use of the system would have on the rights and freedoms of the persons involved, and in particular the seriousness, likelihood and magnitude of those consequences. Not forgetting that, as mentioned above, important systems are excluded from the application of this provision, a further question is which authority will monitor these aspects and on what terms.

As regards the third and fifth paragraphs, Article 5 does not seem to be the right place to recall that "*each use for the purposes of law enforcement of a 're-al-time' remote biometric identification system in publicly accessible spaces shall be subject to a prior authorisation granted by a judicial authority or an independent administrative authority whose decision is binding of the Member State in which the use is to take place, issued upon a reasoned request and in accordance with the detailed rules of national law referred to in paragraph 5*" (art. 5.3) or that, given that these will be measures restricting fundamental rights, States shall detail the conditions under which such authorisations may be requested and obtained, and bring them to the attention of the Commission no later than 30 days after the date of application *for authorisation.* 5.3) or that, given that they will be measures restricting fundamental rights, Member States must detail the conditions under which such authorisations may be requested and obtained, and bring them to the attention of the Commission no later than 30 days after their adoption (art. 5.5), and monitor them and report annually to the Commission.

Today, as has already been repeatedly stated, the constitutional systems of the Member States explicitly require such authorisations -as well as a sound legal basis for them- and also provide for situations in which prior authorisations can be relaxed due to particular urgency. In addition, the data used by these systems are biometric data covered by art. 4.14 GDPR (and affected by Art. 9 GDPR). Reiterating it in article 5 does not detract from the validity of these other legal provisions, but it also does not provide significant assistance.

Really, the only novelty in this respect is the one contained in art. 5.4, which requires that *'each use of a 'real-time' remote biometric identification system in publicly accessible spaces for law enforcement purposes shall be notified to the relevant market surveillance authority and the national data protection authority in accordance with the national rules referred to in paragraph 5'.* This information obligation is interesting, although again it should be recalled that systems used for military, defence or national security purposes employed by Member States, according to art. 2.3 AIA, shall not be part of that whole. Given that, art. 9.2 GDPR leaves little scope for private operators to use such technologies and, more importantly, that art. 5 AIA does not seek to prohibit such uses by private operators, these paragraphs 2 to 8 of art. 5 AIA are, indeed, bordering on irrelevant.

V. Biometric recognition modalities classified as high risk

The regulation of the treatment of high-risk AI systems is an issue to which the AIA devotes a large part of its articles, as discussed in detail in another chapter of this treaty by Professor Cotino Hueso. In addition to the systems listed in Annex III of the Regulation, the list that will be dealt with in this section, it should be borne in mind that the Commission is empowered to update this list by means of delegated acts.

The fact that a biometric recognition system is covered by this set of modalities is of great significance, given the requirements that the AIA establishes with regard to risk management, mechanisms to avoid negative bias, the obligation to have automatic records of system activity, human supervision or, among others, the appropriate level of accuracy, robustness or cybersecurity. It is true that Title IX opens up the possibility for the Commission and Member States to support the drafting of *voluntary Codes of Conduct* for application in non-high risk systems, but there is no scope for comparison between measures that are mandatory in a Regulation and self-monitoring measures. AIA was expected to be a key part of the EU's new digital constitutionalism[124], but it has fallen far short of that expectation.

As mentioned above, on the basis of the wording of Article 6 and, in particular, Annex III of the Regulation, we find that some biometric recognition systems are classified as *high-risk systems* in two subgroups.

Firstly, regardless of the operational scenario in which they are used -point 1 of Annex III- the following are high risk: (a) remote biometric identification systems; (b) biometric categorisation systems based on sensitive attributes or characteristics; and (c) AI systems intended to be used for biometric emotion recognition. In any case, recognition systems offering authentication or verification functionalities are excluded.

A second subset of high-risk modalities is detailed in Annex III, points 2 to 8. In particular, a list of 21 high-risk AI modalities is provided, grouped into six operational scenarios: education and vocational training; employment; essential services and benefits; law enforcement; cross-border transit; and administration of justice. The wording of these high-risk forms of AI has been

[124] De Gregorio, G., "The Rise of Digital Constitutionalism in the European Union", *International Journal of Constitutional Law* 19.1, 41-70, 2021.

formulated with a strong emphasis on the term "*AI systems intended to be used for*" actions such as assessment (risk, performance, learning levels, reliability), tracking, detection of prohibited behaviour, classification or decision-making, and may in some of these cases include: biometric recognition systems, which cannot be understood as falling into the three categories described in the first bullet point, nor into the excluded category of authentication.

The systematic approach used to organise the presentation of high-risk modalities has not been very adequate. Biometric technologies take on roles in operational domains and the term *biometric application* can be used to refer to the combination of functionalities, uses and roles they play[125]. The operational objectives of biometric technologies are as diverse as the reasons why we want to identify people or try to find some of them[126]. Some systems search for people whose biometrics we know (enrolled) and others for people whose biometrics -even identity- we do not know. Some systems work on claimed identities, and others check for the presence of an unclaimed identity, or may not strictly speaking care so much about an individual's identity as about his or her group membership. From this point of view, it is questionable why the legislators have defined the lists in paragraphs 2 to 8 of this Annex III so narrowly in a list of operational scenarios -some of which have not been defined for the purposes of the AIA-.

With respect to this second set, and provided that they fit the description of these operational scenarios and utilities in the annex, one can think of: non-remote biometric identification scenarios; biometric categorisations not based on sensitive attributes; or detections of behavioural biometrics that do not fit, properly speaking, in the category of emotion recognition.

An example in the educational operational scenario (point 3(d)[127]) could be a biometric system which, in the educational scenario, identifies prohibited behaviour during exams by means of a biometric recognition system trained to detect atypical behaviour (e.g., tendency to hide hands under desks, simulation of writing, conspicuous deviations in gaze fixation, etc.). In the workplace, on the other hand, one can think of systems such as those used in behavioural assessments which, in accordance with point 4(b), can be used to assign tasks to individuals, to promote them in their careers or even to lead to the termination of their contracts. However, these will not always be remote

[125] Escajedo San Epifanio, L., *Tecnologías Biométricas*, cit., 2017,244-248.

[126] Wayman, J./ Jain, A. K./ Maltoni, D./ Maio, D., cit., 2005,4-5.

[127] Annex III. 3. d) AI systems intended to be used for the monitoring and detection of prohibited behaviour by students during examinations in the context of or within educational and vocational training institutions at all levels.

or categorising systems, raising questions about the impact on interpretation of the exclusion of non-remote singling out systems.

Also, in relation to Block 5 of Annex III, it may be the case that non-remote identification systems may be employed in respect of situations such as those described for emergency calls (Annex III.5 (d)) or biometric categorisations and behavioural detections other than those described in point 1 of the Annex may be understood to be covered by statements such as those of systems to assess the risks of criminality or recidivism (point 6(d), the reliability of a witness or expert witness (as part of point 6(c), which concerns the assessment of the reliability of evidence) and certainly in the management of border transit where travel documents that can be authenticated by biometric verification are not available (point 7(d)).

The AIA expands the list of supervisory entities for these systems to include not only the national notifying authority and the market supervisory authority, but also those responsible for overseeing security, migration, or asylum activities, as well as data protection agencies. It should not be forgotten in any case that the latter have, in turn, powers assigned to them by the GDPR.

VI. Large-scale recognition systems operational before the entry into force of the Regulation (Art. 111 and Annex X)

Consistent with the expansion of biometric identification to the travel documents of nationals -which the US has not yet done- Member States started to develop information systems with biometric components in the context of the Schengen Agreement, the Prüm Treaty, the visa system[128] or the Dublin Convention, under which EURODAC emerged. At the service of or as part of such information systems, there are currently a number of biometric recognition tools -with a greater or lesser operational level, depending on the case- which are used, inter alia, against illegal immigration, terrorism, or trafficking in human beings. These applications, it has also been said, basically apply one-to-one and one-to-many identifications with the participation of individuals, with very few exceptions (limited to criminal prosecution or, in some cases, to the identification of possible victims). This circumstance, applying the provision of the AIA with regard to verifications (vid. II.3) leaves a large part of these biometric recognition modalities outside its orbit or, at most, awaiting codes of conduct or voluntary guidelines.

In spite of resistance in the past, a number of decisions have been tak-

[128] EC Regulation No 767/2008.

en in recent years which, subject to many data protection requirements, are moving towards the interoperability of these large-scale IT systems, which are strategic in the European area of freedom, security and justice. Throughout the drafting of the AIA, especially with regard to immigration not linked to any type of *criminal conduct*, some voices were in favour of dismantling this type of recognition systems or, at the very least, excluding them from this interoperability macro-project. For many reasons, however, the legislative process could not stop at reviewing the complex reality of these systems.

This led to the conclusion reached in Article 111 AIA in connection with its Annex X. According to these rules, large-scale IT systems already introduced or planned to be introduced before 36 months after the entry into force of the AIA[129] are in a kind of legal fiction which means that they are considered to be "prior" to the AIA and therefore exempt from its application until 2030, unless there is a substantial reform of any of its elements in the future. It should be noted that this last paragraph is sufficiently ambiguous that it is not really clear whether the AIA will end up applying to these large-scale IT systems or not.

However, this is not the only doubt about the real applicability of AIA to these systems. Delayed searches, as well as one-to-one and one-to-many identifications are subject to notable exclusions in the applicability of AIA, without it being relevant that they handle biometric data falling within the notion of 4.14 of the GDPR and consistently protected as special categories of personal data (9.1 GDPR).

1. Art. 111 and Annex X of Regulation

Article 111.1 provides that, without prejudice to the application of Article 5 in accordance with Article 113[130], paragraph 3(a), AI systems, which are components of large-scale IT systems established under the legislative acts listed in Annex X and which have been placed on the market or put into service before 36 months after the date of entry into force of the AIA -date to be determined- shall be brought into conformity with this Regulation by 31 December 2030 at the latest. The possibility of applying Article 5, however,

[129] See below.

[130] Article 113, Entry into force and application, establishes that the Regulation shall enter into force 20 days after its publication in the Official Journal of the EU, although it establishes a staggered entry into force for several sections of the articles. Thus, the entry into force of Chapters II (prohibitions) and III (high-risk AI) is foreseen 6 months after entry into force, although, as has been seen, the recitals recognise that until such time as the derived acts are adopted, such entry into force will be limited.

may encounter significant difficulties in the light of a systematic interpretation that includes Article 2.3 on the scope of application and the extensive list of exceptions that, as we have seen, can be applied, for example, to cases such as those set out in Article 5.1(h). The time limits are somewhat longer than those allowed for other AI systems which, without prejudice to the prohibitions of Article 5, are already operational when the AIA enters into force.

Returning to large-scale IT systems, starting from January 1st 2031, a conformity assessment of these IT systems is therefore foreseen, either in accordance with the provisions of the legal acts currently governing them, or in accordance with the legal acts that will amend or replace them in the future.

For the purposes of the application of Article 111 AIA, Annex X contains, grouped in 7 blocks, a list of legislative acts, including the Interoperability Regulations and the Databases linked by these regulations, which are dealt with in sections VI.2 and VI.3 below.

2. The Interoperability Regulations

In May 2019, a major interoperability initiative was adopted in the EU -in two Regulations[131], one on borders and visas, and the other on police and judicial cooperation, asylum and immigration.

The police and migration control information systems available in the EU were created at different times and in response to different initiatives, resulting in a fragmented architecture -Commission's expression- where information is not only stored separately but in disconnected forms that facilitate the creation of blind spots[132].

The idea of working towards the interoperability of these systems is not new. As early as 2004, the European Council invited the Commission to make proposals on the interoperability of the EURODAC (and VIS) system, together with other databases, *in order to make this information available for the prevention and combating of terrorism*[133]. Such a strategy, however, was complex

[131] Regulation (EU) 2019/817 of the European Parliament and of the Council of 20 May 2019 establishing a framework for interoperability between EU information systems in the field of borders and visas (OJ L 135, 22.5.2019, p. 27), and Regulation (EU) 2019/818 of the European Parliament and of the Council of 20 May 2019 on establishing a framework for interoperability between EU information systems in the field of police and judicial cooperation, asylum and migration (OJ L 135, 22.5.2019, p. 85).

[132] Leese, M. "Fixing State Vision: Interoperability, Biometrics, and Identity Management in the EU", *Geopolitics*, 27(1), 2020, 113-133.

[133] DE HERT, P./ GUTWIRTH, S.: "Interoperability of Police Databases within the

to promote given the political divergences and differences in the technical capacity of the different Member States. In contrast to the approaches used by countries such as Israel, Canada or the US, especially after the 9/11 attacks[134], in the EU, the EDPS and WP 29 were forceful in expressing their concerns about mass storage in cases such as visas[135], passports or *laissez-passers* and stated that the creation of a centralised database containing personal data and, in particular, biometric data of all persons authorised to receive a passport[136], a visa or a *laissez-passer* was not justified. They considered that this violated the principle of proportionality[137].

With the EU establishing a huge database in Tallinn, Estonia, in 2022, managed by the eu-LISA Agency[138], to collect biometric fingerprint and facial information of more than 400 million individuals from third countries, it is clear that the perspective of Member States and European institutions has changed dramatically.

The path towards interoperability started to become clearer following the recommendations of the High Level Expert Group on Information Systems and Interoperability, set up in 2016 by the Directorate-General for Migration and Home Affairs[139]. The final milestone, however, came with the adoption of the two Interoperability Regulations. The EU definitively embraced a new paradigm in the treatment of biometric recognition for law enforcement and migration control purposes[140]. The criterion of data purpose limitation is now interpreted in a more flexible way than in other areas, giving up the important role of guarantee that it had played since the end of the 1990s[141].

EU?", *International Review of Law Computers & Technology*, vol 20, 1-2/ 2006,21-25; J. A. LEWIS, J. A.: "Biometrics and Security", *Center for Strategic & International Studies*.

[134] Escajedo San-Epifanio, L., *Biometric Technologies*, cit., 2017, 258-261.

[135] Opinion 3/2005 of WG-WP 29 on the SIS II system; Opinion 7/2004 on the inclusion of biometric elements in residence permits and visas taking into account the establishment of the European Visa Information System (WP 96), adopted on 11 August 2004.

[136] Opinion of 23 March 2005 on the proposal for a Regulation of the European Parliament and of the Council concerning the Visa Information System (VIS); Opinion of 19 October 2005 on three proposals concerning the second generation Schengen Information System (SIS II), COM 2005 230 final, COM 2005 236 final, and COM 2005 237 final, OJEU C 91.

[137] C 313/ 38, paragraph 12(3).

[138] European Agency for the Operational Management of Large-scale IT Systems in the Area of Freedom, Security and Justice.

[139] Directorate-General for Migration and Home Affairs, *High-level expert group on information systems and interoperability: Final report,* 2017.

[140] Oliveira Martins, B., Lidén, K./ Jumbert, M. G. "Border security and the digitalisation of sovereignty: insights from EU borderwork". *European Security*, 31(3), 2022, 475-494.

[141] Hartmut A. "Interoperability Between EU Policing and Migration Databases: Risks for Privacy", *European Public Law*, 2020, 26 (1), 93-108.

The interoperability strategy aims to improve the ability of information systems to exchange data, but does not imply that all data are pooled. On a selective basis, and based on the different levels of access of users (such as police, migration officials and border guards) it aims to provide faster, smoother and more systematic access to the information they need to do their work, while ensuring respect for fundamental rights.

From a technical point of view, the key component of this strategy is the creation of a single portal, allowing a single search across all interoperating systems and receiving all available results together. As far as biometric recognition is concerned, interoperability includes a matching service for biometric data obtained from fingerprints and facial images. This biometric matching service is known by the acronym sBMS (*EU shared Biometric Matching System*) and once activated it will be one of the largest biometric recognition systems in the world -second only to India's Aadhaar[142]-. The database will integrate, as already mentioned, biometric fingerprint and facial data of more than 400 million third country nationals[143], for the time being, however, no data of Member States' nationals. Among the processes available will be the extraction of biometric templates from different EU databases, with the aim of simplifying the search and cross-comparison of biometric data. A large and complex system of agencies (e.g., Interpol, Europol), and numerous databases (such as EES, ECRIS-TCN, VIS, EURODAC and ETIAS, described *below*) will be the reference in this interoperable infrastructure for migration surveillance and crime control in the EU[144].

There will be a common repository of identities, into which the biographical information of non-EU citizens already available in the databases covered by the strategy will be incorporated, but it should be noted that, in principle, the Interoperability Regulations are not a legal basis for adding to the information already available. The legitimate basis for storing, adding,

[142] *Aadhaar* is currently considered the largest biometric identification system in the world, created by the government with the aim of incorporating the data of all persons residing in India, regardless of their citizenship, and it is estimated that it has currently enrolled more than 1.2 billion people. Vid. Escajedo San-Epifanio, L., *Tecnologías Biométricas*, cit., 2017, 275-276; Kloppenburg, S./ Van der Ploeg, I., "Securing Identities: Biometric Technologies and the Enactment of Human Bodily Differences", *Science as Culture*, 29(1), 2018, 57-76.

[143] Jones, C., "Data protection, immigration enforcement and fundamental rights: what the EU's regulations on interoperability mean for people with irregular status", *Statewatch and Platform for International Cooperation on Undocumented Migrants*, 2019, 6.

[144] Oliveira Martins, B., Liden, K./ Jumbert, M. G. "Border security and the digitalisation of sovereignty: insights from EU borderwork", *European Security*, 31(3), 2022, 475-494.

modifying or deleting data in each of the databases will remain the legislative act regulating each of them.

A multiple identity detection mechanism and a set of data quality control mechanisms are also promoted, and a number of budget lines are foreseen, both for eu-LISA, Europol, CEPOL and the European Border and Coast Guard Agency, and for Member States to equip themselves with technical components and training to enable their officers to participate in the user community, although not all Member States participate under the same conditions in the Schengen arrangements[145].

3. Biometric recognition systems covered by Annex X: some relevant data

At the time of the adoption of the Interoperability Regulations, three of the six systems expected to be covered by the strategy were in place: SIS, Eurodac and VIS.

The Schengen Information System (SIS)[146] contains a wide range of alerts on persons (refusals of entry or stay, EU arrest warrants, missing persons, assistance in judicial proceedings, discreet checks) and objects (including lost, stolen or invalidated identity or travel documents).

The EURODAC system[147] contains the fingerprint data of asylum seekers and third-country nationals who have irregularly crossed the external borders or who are staying illegally in a Member State. EURODAC was the first institutionally based automated biometric recognition system in the EU, and

[145] The Schengen Agreement is an agreement whereby several countries in Europe abolished internal border controls (between these countries) and moved these controls to external borders (with third countries). The following countries are currently part of the Schengen area: Austria, Belgium, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Italy, Latvia, Liechtenstein, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Romania, Slovak Republic, Slovenia, Spain, Sweden and Switzerland.

[146] Regulation (EU) 2018/1860 on the use of the Schengen Information System for returning illegally staying third-country nationals; Regulation (EU) 2018/1861 on the establishment, operation and use of the Schengen Information System (SIS) in the field of border checks; Regulation (EU) 2018/1862 on the establishment, operation and use of the Schengen Information System (SIS) in the field of police cooperation and judicial cooperation in criminal matters.

[147] Amended proposal for a Regulation concerning the establishment of "Eurodac" for the comparison of biometrics for the effective implementation of two future Regulations, the Regulation on asylum and migration management and the Regulation on resettlement, and amending Regulations (EU) 2018/1240 and (EU) 2019/818 – COM(2020) 614 final.

became operational in January 2003, based on Regulation (EU) 2752/2000, for the implementation of the Dublin Convention[148]. It is a checking system that verifies in a centralised database whether or not particular fingerprints are registered[149] and are found to belong to someone who has applied for asylum in another country, someone who has been refused asylum or someone who, for some reason, is not allowed to apply for asylum. This is to avoid *asylum shopping*, i.e., the simultaneous application for asylum in several countries with the aim of choosing the most favourable one, or attempts to gain access under different identities[150]. However, its use has expanded to cover not only those who actually apply for asylum, but all *potentially* irregular migrants of increasingly young age.

The third and last of the systems prior to the Interoperability Regulations is the Visa Information System (VIS)[151], which operates with data relating to holders of short-stay visas. The VIS is a system created to support the issuance of a type of visa that is valid for the whole EU and replaces those previously issued by States. Regulation EC No 767/2008 on the Visa Information System foresaw a progressive implementation of the VIS, according to criteria such as the risk of illegal immigration, threats to the internal security of Member States and the feasibility of collecting biometric data from all locations in a region. This visa information system was launched in 2009 for some regions in the world, in particular with a view to those with the

[148] The Dublin Convention was replaced by Council Regulation (EC) No 343/2003.

[149] Van der Ploeg, I./ Sprenkels, I. "Migration and the Machine-Readable Body", in Van der Ploeg/ Sprenkels (eds), Migration and the New Technological Borders of Europe, Springer, 2011, 83-84.

[150] On 1 June 2013, eu-LISA took over the day-to-day operational management of EURODAC from the Commission. The central server is a fully automated system. In 2015, the new EURODAC Regulation (603/2013) entered into force, allowing national police forces and EUROPOL to access the database for the prevention, investigation and detection of criminal activities. On 4 May 2016, the European Commission proposed (COM 2016/0132 COD) to strengthen and extend the EURODAC Regulation, and in 2018, in a provisional agreement, the Parliament and Council agreed on an extension of the system. As part of the broader migration and asylum pact, the Commission presented an amended proposal on 23 September 2020 (COM (2020) 614). If accepted, the proposal would introduce an obligation to store data on names, nationalities, place and date of birth, and information on travel documents; for asylum seekers, the obligation is to store the asylum application number and the Member State responsible under the Dublin Regulation.

[151] Visa Information System: Proposal for a Regulation amending Regulation (EC) No 767/2008, Regulation (EC) No 810/2009, Regulation (EU) 2017/2226, Regulation (EU) 2016/399, Regulation XX/2018 [Interoperability Regulation] and Decision 2004/512/EC, and repealing Council Decision 2008/633/JHA – COM (2018) 302 final.

most difficulties in providing indubitable travel documents to their nationals. North Africa, the Middle East and the Gulf region were selected.

When the Interoperability Regulations were drawn up, three additional systems were in preparation, two of which -since their design- had a high degree of interoperability both with each other and with the VIS. They are the following:

1. An Entry-Exit System (EES)[152], which has been adopted and will replace the current manual passport stamping system, will electronically record the name, type of travel document, biometric data, date and place of entry and exit of third-country nationals visiting the Schengen area for a short stay.

2. A European Travel Information and Authorisation System (ETIAS)[153], which, once adopted, will be a largely automated system that will collect and verify security-related information submitted by visa exempt third country nationals prior to their travel to the Schengen area.

3. And a European Criminal Records Information System for third country nationals (ECRIS-TCN system)[154], which, once adopted, would be an electronic system for the exchange of information on previous convictions of third country nationals by EU criminal courts.

In addition to these databases, the interoperability strategy also foresaw to include links to Interpol's Stolen and Lost Travel Documents (SLTD) database, which should be systematically consulted at the EU's external borders, and Europol's data. Interoperability with national information systems and EU decentralised information systems was not foreseen.

[152] Regulation (EU) 2017/2226 establishing an Entry-Exit System (EES) for recording entry and exit data and refusal of entry data of third-country nationals crossing the external borders of the Member States, determining the conditions of access to the EES for law enforcement purposes and amending the Convention implementing the Schengen Agreement and Regulations (EC) No 767/2008 and (EU) No 1077/2011.

[153] European Travel Information and Authorisation System. Regulation (EU) 2018/1240 of the European Parliament and of the Council of 12 September 2018 establishing a European Travel Information and Authorisation System (ETAIS) and amending Regulations (EU) No 1077/2011, (EU) No 515/2014, (EU) 2016/399, (EU) 2016/1624 and (EU) 2017/2226 (OJ L 236, 19.9.2018, p. 1).... Regulation (EU) 2018/1241 of the European Parliament and of the Council of 12 September 2018 amending Regulation (EU) 2016/794 in order to establish the European Travel Information and Authorisation System (ETAIS) (OJ L 236, 19.9.2018, p. 72).

[154] European Criminal Records Information System on third-country nationals and stateless persons. Regulation (EU) 2019/816 of the European Parliament and of the Council of 17 April 2019 establishing a centralised system for the identification of Member States holding information on convictions of third-country nationals and stateless persons (ECRIS-TCN) to complement the European Criminal Records Information System and amending Regulation (EU) 2018/1726 (OJ L 135, 22.5.2019, p. 1).

The legislative and operational work towards interoperability, as well as the funds used for this purpose, are proof of the strategic importance that the EU attaches to these recognition tools. Indeed, in addition to the foundations relating to cross-border transit, there has been determined progress in police cooperation for the prosecution of serious crime. At the end of 2023, for example, an agreement was reached on automated data exchange for police cooperation under the Prüm Treaty, which allows law enforcement authorities to consult the national databases of other member states for DNA, fingerprint and vehicle registration data. The new Prüm Regulation, which will reflect the agreements reached, will involve the installation of a router by eu-LISA (EU agency in charge of large-scale IT systems, such as the Schengen Information System) to facilitate the establishment of connections between member states (and EUROPOL) to retrieve data. The router will consist of a search tool and a secure communication channel, and will forward the query request submitted in one Member State to all Member States and EUROPOL.

Among the major criticisms of interoperability are those that refuse to interlink, even through the possibility of cross-checking, migration control and the prosecution of the most serious crimes[155], mixing -at a point of no return- databases originally created for very different purposes[156]. It has also been criticised that the EU finances, in the countries of origin, the development of the transnational WAPIS programme. This is a system managed by INTERPOL to enhance the capacity of law enforcement agencies in West Africa to combat transnational organised crime and terrorism through cross-border exchange. Its basis, however, is a digitisation of biometric data of African citizens into formats that are subsequently readable with larger international databases and organisations, such as FRONTEX. In view of this, critics speak of a 'deterritorialisation' of the EU's external borders[157], which are somehow transplanted to certain African countries. Since 2017, WAPIS

[155] Vavuola, N., "The recast Eurodac regulation: are asylum seekers treated as suspected "crimminals"?", in C. Bauloz et al., eds. *Seeking asylum in the European Union: selected protection issues raised by the second phase of the common European asylum system*, Brill, 2015, 247-273; Queiroz, M.B., 2019. The impact of EURODAC in European migration law: the era of crimmigration? Market and competition Law review, 3 (1), 157-183.

[156] Bunyan, T., "The "point of no return" interoperability morphs into the creation of a Big Brother centralized EU state database including all existing and future", *Justice and Home Affairs databases. Statewatch Analysis*, 2018.

[157] Oliveira Martins, B/ e. al., "Border security", cit., *European Security*, 2022, 31(3), 475-494.

has been applied in all member states of the Economic Community of West African States[158], founded by the Treaty of Lagos, and in Mauritania.

VII. Final reflections

At the time of the legislative acts that introduced the biometric Passport in the EU, Professor Stefano Rodotá warned that the human body *"had become a password*". It was a time when personal data protection was beginning to be characterised by strong contradictions, reflecting a "*real social, political and* institutional *schizophrenia*"[159].

Recognition technologies should have started to be tested not so much for the needs they could fulfil, but for the real admissibility of the concrete uses they were intended to be put to. At the time, after the 9/11 attacks, it did not seem a good time to reflect on how much space we wanted to give to biometric recognition technologies, but it was not foreseen that they would end up being used for such banal purposes as opening a locker in a gymnasium. The mechanisms for guaranteeing fundamental rights were not prepared to deal with the uniqueness of biometric information and experience has shown that, in more than a few cases, it has been necessary to review important judicial decisions. For example, just over a year ago, Spain had to review the doctrine that the Supreme Court had established regarding the biometric control of workers' attendance. In July 2007,[160] issued a ruling in which it did not seem to see a major problem with manual biometric clocking in. In its 7th legal basis, the SC went so far as to say, "*it seems as if the trade unions that have promoted the process see a binary code of the three-dimensional image of the hand as an affront to human dignity. But the scope of the system does not go that far*". The unification of criteria promoted by the European Data Protection Committee established in April 2023, establishes that, given that biometric identification data belong to the special category of art. 9 GDPR, they can only be processed in

[158] Members are Benin, Burkina Faso, Cape Verde, Côte d'Ivoire, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Mali (suspended from 2021), Niger (suspended from 2023), Nigeria, Senegal, Sierra Leone and Togo.

[159] In the same vein, N. P. MUSAR, in AA. VV., "Workshop report. Restrictions on the Implementation of the EU Data Protection Directive for Public Interest, security and defence", in *HIDE Newsletter*, vol. 2 (7), 2009, December, 2 ff; HARB, B./ SCHMID, D.: "Der Einsatz biometrischer Systeme. Verfassungsrechtlische Aspekte", cit. 2005, 158-161.

[160] Judgment of 2 July 2007, heard by the Administrative Chamber of the Supreme Court, appeal no. 5017/2003, on the implementation of the new time control system.

cases in which a measure of legal rank specifically enables their use, as well as the corresponding guarantees.

Most probably, in a decade or sooner, we will also need to review the extent to which the AIA allows the handling of non-personally identifiable biometric data by private operators -especially in the field of commerce. Professor Francisco Balaguer Callejón warns *that* the digital world, which occupies an increasingly important part of our daily reality, "*is subject to rules in the production of which the State practically does not intervene and which do not conform to* constitutional *principles and values*"[161]. And in the case of biometric surveillance, we find a worrying trend towards human beings becoming detectable, traceable, and correlatable without their knowledge or consent and for a wide variety of purposes[162]. Allowing the uncontrolled collection of bodily data falls far short of the minimisation principle.

As a final thought, and as far as automated biometric recognition is concerned, the AIA will not go down in history for its great contributions. So much text, to say so little and so badly. As soon as it appears to regulate a particular case, it immediately becomes entangled in so many nuances and qualifications that it is very difficult to determine the real scope of its precepts.

The verification modalities, although they use special category personal data, fall outside the AIA without any justification, and it seems that non-remote biometric identification techniques in which the subject to be identified participates actively, whether voluntarily or not, have been included with them. The fact that in both cases they handle special category biometric data (Art. 4.14 and Art. 9 GDPR) has not been sufficient reason to provide these systems with the quality and security analyses that, in accordance with the AIA, will be applied to high-risk modalities, leaving them, at most, pending what may be proposed in the Voluntary Guidelines.

Even the most tangible, high-risk prohibitions and modalities, are gradually emptied of content as one reads the text finally adopted. The most important exclusion implies that none of these modalities apply to the military, defence and national security uses of the Member States (art.2.3 AIA)[163]; not even when it is private actors who provide this service to the Member States.

[161] Balaguer Callejón, F., *La Constitución del Algoritmo*, 2nd ed., Fundación Manuel Giménez Abad, Zaragoza, 2023, 33.

[162] Gutwith, S. / De Hert, P., "Regulation Profiling in a Democratic Constitutional State", in M. Hildebrandt/ S. Gutwith (eds.) *Profiling the European Citizen*, 2008, 287.

[163] If and to the extent that AI systems are placed on the market, put into service or used, with or without modification, for military, defence or national security purposes, they should be excluded from the scope of this Regulation, irrespective of the type of entity carrying out such activities, for example irrespective of whether they are a public or a private entity.

Only some modalities applicable in the employment or education sectors, but not many, seem to find ways of prohibition or limitation, because even the commercial sector has found an escape route in the AIA.

As if this were not enough, we should be astonished by the ease with which categorisation or emotion recognition systems are admitted for use in commercial environments under the argument that this type of practice -especially when it does not use art. 4.14 data- does not generate as much risk for the rights of individuals as one might think. This formulation, which speaks of a lower risk, but not of an absence of risk, does not serve as an excuse to develop the necessary safeguards.

The process of drafting the AIA, constrained by the format of a product risk regulation, was not the right time to regulate issues relating to automated biometric recognition, because, as we have seen, more careful thought was needed. And its fruit, the text finally adopted, reflects this. It is paradoxical -and rather sad- that it is precisely the allusion to biometrics in the AIA that appears in so many speeches as a sign of a high commitment to the values of the Union. It would have been difficult to do worse.

Given their high level of invasiveness, a clear regulation of biometric recognition systems, solid in its formulation and with tangible and efficient guarantees, is necessary and essential. As a first step, it will be necessary to review both the AIA, in practically all its provisions on biometric recognition, and some of the assumptions made in the GDPR.

THE PROHIBITION OF ARTIFICIAL INTELLIGENCE SYSTEMS THAT EVALUATE AND CLASSIFY PEOPLE BASED ON DATA THAT ARE UNRELATED TO THE CONTEXT IN WHICH THEY WERE GENERATED AND THAT LEAD TO DISCRIMINATION

Miguel Ángel Presno Linera[1]
Professor of Constitutional Law at the University of Oviedo

I. Introduction

The AIA includes a Recital 31 which explains that "AI systems providing social scoring of natural persons by public or private actors may lead to discriminatory outcomes and the exclusion of certain groups. They may violate the right to dignity and non-discrimination and the values of equality and justice. Such AI systems evaluate or classify natural persons or groups thereof on the basis of multiple data points related to their social behaviour in multiple contexts or known, inferred or predicted personal or personality characteristics over certain periods of time. The social score obtained from such AI systems may lead to the detrimental or unfavourable treatment of natural persons or whole groups thereof in social contexts, which are unrelated to the context in which the data was originally generated or collected or to a detrimental treatment that is disproportionate or unjustified to the gravity of their social behaviour. AI systems entailing such unacceptable scoring practices and leading to such detrimental or unfavourable outcomes should therefore be prohibited".

In other words, we are talking about the fact that after the entry into force of the Regulation, AI systems that generate qualifications or social hierarchies of people based on their behaviour or characteristics and that may give rise to discriminatory situations and, therefore, violate the principles of dignity and equality, will be prohibited in the European Union and may not be exported to other countries.

This is an extremely important issue because essential elements of the social and democratic rule of law are at stake, which would be seriously undermined if systems, such as those mentioned above, aimed at conditioning the behaviour of citizens and capable of generating, at the very least,

[1] This work is one of the results of Project PID2022-136548NB-I00 "Los retos de la inteligencia artificial para el Estado social y democrático de Derecho", funded by the Ministry of Science and Innovation in the Convocatoria Proyectos de Generación de Conocimiento 2022.

social and economic damage, if not physical and psychological harm, were allowed[2].

In the following pages we will develop the hypothesis of the certain risk posed by the systems we are dealing with[3] and the success, in our opinion, of their introduction in the European Regulation as a way of dealing with one of the growing manifestations of, in Shoshana Zuboff's words, "surveillance capitalism"[4].

Obviously, it is not a matter of excluding any type of personal "punctuation" that aims at behavioural modifications, since there are systems that are not only possible but surely necessary; for example, the points-based driving licence would be a good and well-known example: in the words of the Directorate General of Traffic, its objective "is to modify the behaviour and attitudes of offending drivers, to make them aware of the serious human, economic and social consequences of traffic accidents and to make them see the implication of their behaviour in accidents"[5].

In other contexts, the systems will not be prohibited but will be classified as "high risk"; thus, "AI systems used in employment, workers management and access to self-employment, in particular for the recruitment and selection of persons, for making decisions affecting terms of the work-related

[2] See Paquale, F. and Keats Citron, D., "The Scored Society: Due Process for Automated Predictions", *Washington Law Review*, vol. 89, 2014, pp. 1-33.

[3] On algorithmic risk San Martín Segura, D., *La intrusión jurídica del riesgo*, CEPC, Madrid, 2023, pp. 271 et seq.

[4] "Surveillance capitalism, m. 1. A new economic order that claims human experience as a free raw material to be exploited for a series of hidden commercial practices of extraction, prediction and sales. 2. Parasitic economic logic in which the production of goods and services is subordinated to a new global architecture of behavioural modification. 3. Unscrupulous mutation of capitalism characterised by vast concentrations of wealth, knowledge and power that are unprecedented in human history. 4. The fundamental framework for a surveillance economy. 5. As great a threat to human nature in the twenty-first century as industrial capitalism was to the natural world in the nineteenth and twentieth centuries. 6. The origin of a new instrumental power that imposes its domination on society and poses alarming contradictions for market democracy. 7. A movement that aspires to impose a new collective order based on absolute certainty. 8. Expropriation of crucial human rights that can perfectly well be considered a coup from above: an overthrow of the sovereignty of the people", *La era del capitalismo de vigilancia*, Paidós, Barcelona, 2022, 2nd edition, p. 9.

[5] "The points balance can change: being a good driver and/or taking awareness courses earns you points. Committing offences subtracts points, until you reach zero. If you reach this point, your licence will be revoked and you will not be able to drive any vehicle, although before this happens, you can recover points", available at https://www.dgt.es/nuestros-servicios/permisos-de-conducir/tus-puntos-y-tus-permisos/como-funciona-el-permiso-por-puntos/ (as of 18 March 2024).

relationship, promotion and termination of work-related contractual relationships, for allocating tasks on the basis of individual behaviour, personal traits or characteristics and for monitoring or evaluation of persons in work-related contractual relationships, should also be classified as high-risk, since those systems may have an appreciable impact on future career prospects, livelihoods of those persons and workers' rights. Relevant work-related contractual relationships should, in a meaningful manner, involve employees and persons providing services through platforms as referred to in the Commission Work Programme 2021. Throughout the recruitment process and in the evaluation, promotion, or retention of persons in work-related contractual relationships, such systems may perpetuate historical patterns of discrimination, for example against women, certain age groups, persons with disabilities, or persons of certain racial or ethnic origins or sexual orientation. AI systems used to monitor the performance and behaviour of such persons may also undermine their fundamental rights to data protection and privacy" (Recital 57 of the European Regulation).

Similarly, "access to and enjoyment of certain essential private and public services and benefits necessary for people to fully participate in society or to improve one's standard of living. In particular, natural persons applying for or receiving essential public assistance benefits and services from public authorities namely healthcare services, social security benefits, social services providing protection in cases such as maternity, illness, industrial accidents, dependency or old age and loss of employment and social and housing assistance, are typically dependent on those benefits and services and in a vulnerable position in relation to the responsible authorities. If AI systems are used for determining whether such benefits and services should be granted, denied, reduced, revoked or reclaimed by authorities, including whether beneficiaries are legitimately entitled to such benefits or services, those systems may have a significant impact on persons' livelihood and may infringe their fundamental rights, such as the right to social protection, non-discrimination, human dignity or an effective remedy and should therefore be classified as high-risk. " (Recital 58).

As is well known, and as explained in more detail in other sections of this collective work, the classification of a system as high risk implies a series of obligations; among others:

"High-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers." (Article 13.2);

> "High-risk AI systems shall be designed and developed in such a way, including with appropriate human-machine interface tools, that they can be effectively overseen by natural persons during the period in which they are in use. 2. Human oversight shall aim to prevent or minimise the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse..." (Article 14.1 and 2);
>
> "High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle." (Article 15.1);
>
> "... High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures" (Article 15.4.3).

II. The Chinese social credit system

The most talked-about social scoring system, which began to develop even before the current AI boom, is the Chinese social credit system (hereafter CSCS); as Lauren Yu-Hsin Lin and Curtis J. Milhaupt[6] explain, planning for a comprehensive social credit programme to complement China's weak legal system began in the 1990s with the more ambitious goal of addressing widespread fraud in the country's transition from central planning to a fledgling market economy. Those efforts culminated in 2014 with the joint publication by the Chinese Communist Party Central Committee and the Chinese State Council of the *Planning Outline for Building a Social Credit System (2014-2020)*, a comprehensive programme to assess the social credit of individuals, enterprises, government entities and other organisations.

Today, the social credit system is also the centrepiece of China's digital governance strategy, marking a shift towards a self-regulating market, i.e., one

[6] "China's Corporate Social Credit System: The Dawn of Surveillance State Capitalism?", *The China Quarterly*, Cambridge University Press, 2023, pp. 1-19; in particular, pp.2-4; available, as of 18 March 2024, at https://www.cambridge.org/core/journals/china-quarterly/article/chinas-corporate-social-credit-system-the-dawn-of-surveillance-state-capitalism/EC80AC0CC9AE60D3D3C631A707A5CE54 (as of 18 February 2024); see also Rogier CREEMERS "China's Social Credit System: An Evolving Practice of Control", 9 May 2018, available, as of 18 February 2024, at https://ssrn.com/abstract=3175792 and http://dx.doi.org/10.2139/ssrn.3175792 (as of 18 March 2024).

in which actors are pressured or incentivised to conform their behaviour to party-State norms beyond the ordinary channels of law and regulation.

In the private sphere, Alibaba introduced its own personal credit scoring system*, Sesame Credit*, as early as 2015, to collect information on personal identity, credit history, contractual reliability, and social behaviours and relationships. Based on this information, participants are assigned social credit scores that are visible to others, and those with high scores are offered advantages, such as faster loan approval[7].

Zuboff explains that the *Sesame Credit* system generates a "holistic" assessment of a person's "character" through algorithmic learning that assimilates much more than whether they pay their bills and loans on time. Algorithms evaluate and rank purchases (for example, whether they are video games rather than children's books), educational titles, and things like the quantity and "quality" of friendships. Well-scored individuals receive distinctions and rewards from *Sesame Credit* customers in their behavioural futures markets. They

[7] On its website *Sesame Credit* explains: "The concept of a credit score may feel complicated, but in essence it looks simply at your payment history, amount of debt, how long you have had debt and how many recent applications you have made for credit accounts. Information about these items are reported to the three credit bureaus, Experian, TransUnion and Equifax, who compile your credit report. The information on your credit report is used to calculate your credit score. Your three-digit credit score captures your experiences with credit and debt and can help you track changes in your financial history over time, from the very first debt you encounter-such as the credit card you opened in college-up to the present. Credit score is a powerful tool that signals to prospective lenders your ability to make payments in a timely manner. This number is unique to you but publicly available under federal law to lenders considering you as a borrower. Your score can be a point of personal pride for good financial management and a point of public documentation. A credit score is an easy way to explain to another person or prospective lender that you can honor your commitment to make timely payments on outstanding debts. In turn, higher scores might lead a lender to extend interest rates lower than they would for consumers with less-favorable credit scores. You can get your credit score as part of a request for a credit report or independently of a credit report. A comprehensive solution is to open a free Credit Sesame account. This provides you with fast access to everything you need to know about your credit history, including your credit score. It includes helpful supporting information that makes sense of your score and report....

Legally, a variety of entities and people can request a copy of your creditreport, which is the information that feeds into your credit score. According to the Consumer Financial Protection.

Bureau (CFPB), this list includes: Businesses to whom you owe money, Government agencies.

Landlords, Employers, Insurance providers, Banks and financial providers, Legal entities (in the event of court orders, for example), Others you have authorised in writing to receive a copy"; available, as of 18 March 2024, https://www.creditsesame.com/knowledge-hub/what-is-credit-score/.

may be able to rent a car without paying a deposit, or receive more favourable terms on that loan or flat rental they apply for, or have their visa application expedited, or receive more prominent exposure on dating applications, and so on. However, some testimonies suggest that the privileges associated with a high personal reputation can suddenly turn into penalties for reasons completely unrelated to a person's behaviour in their role as a consumer: for example, if they have cheated on an exam at university[8].

Turning to the CSCS, it has two main features: the first is the collection of nationwide data from a wide range of regulatory bodies, central and local governments, the judiciary and private platforms. When fully operational, the system will collect two basic types of information: public credit information, generated by a company's interactions with government bodies and regulatory agencies (fines, judgements, business licences...), and market credit information, generated by a company's interactions with other market players (consumer complaints, data generated by credit rating agencies...). The data will be used in scoring systems run by local administrations, most of which are under construction.

The second main element of the CSCS is a regime of rewards and punishments (in the form of "red lists" and "black lists") maintained by government agencies. Some lists have a broad scope, such as non-compliance with court rulings, while others apply to specific sectors of the economy, such as food or medicine.

The inclusion in a red or black list is public; in the former case, it may entail various benefits, ranging from increased access to loans to reduced frequency of inspections or increased opportunities in public procurement processes and access to funding, especially for small and medium-sized entities. Blacklisting creates market barriers, such as restrictions on obtaining government approvals, increased frequency of inspections and prohibitions on obtaining funding. When an entity is blacklisted, its legal representative and the persons directly responsible for the violation will also be blacklisted[9].

[8] *Ob. cit.*, pp. 520 and 521.

[9] Yu-Hsin Lin and Curtis J. Milhaupt, ob. cit., pp. 3-4; more extensively, Schaffer, K., "China's social credit system: context, competition, technology and geopolitics." *Trivium China*, 16 November 2020, available, as of 18 March 2024, at https://www.uscc.gov/sites/default/files/2020-12/Chinas_Corporate_Social_Credit_System.pdf See also Lam T. "The People's Algorithms: Social Credits and the Rise of China's Big (Br)other", in Mennicken, A. Salais, R. (eds) *The New Politics of Numbers. Executive Politics and Governance*, Palgrave Macmillan, 2022; pp. 71-95; especially pp. 78 ff; Xu XU, Kostka, G. and Cao, X. "Information Control and Public Support for Social Credit Systems in China", The Journal of Politics, Vol. 84, no. 4, 2022, pp. 2231-2245, https://www.journals.uchicago.edu/doi/10.1086/718358 (as of 18 March 2024).

III. The development of qualification systems as a way of expanding surveillance capitalism

In a note at the beginning of these pages we collected the definitions of "surveillance capitalism" proposed by Zuboff, and the first two meanings, with some qualifications, seem to encompass practices such as those that characterise the Chinese social credit system: they would be, firstly, part of a new economic-political order that claims for itself human experience as free raw material exploitable for a series of hidden political, social, and commercial practices of extraction, prediction, and sales; secondly, they would be presided over by a parasitic logic in which the production of goods and services is subordinated to a new global architecture of behavioural modification.

It does not appear that the AIA's provision prohibiting AI systems that provide social ratings of natural persons for general use is intended to address the implementation or use in Europe of systems such as the Chinese social credit system: in EU countries and other democratic states, privacy and personal data enjoy a high level of legal protection and there is a higher degree of social concern about the threats that tools of this nature pose to these rights and to the free development of individual personality; as a result, practices typical of totalitarian societies have not developed, such as the so-called "*dang'an*, the personal file of multiple and varied aspects of each of the hundreds of millions of urban inhabitants that is updated from their childhood and for the rest of their lives. This "Mao-era system for recording the most intimate details of life" draws on up-to-date information provided by teachers, Communist Party officials and employers. Citizens have no right to check the contents of their own files, let alone challenge them"[10].

Notwithstanding these differences between the European and Chinese "ecosystems", it should be noted for the sake of nuance that, firstly, the so-called "privacy paradox" is present here: while individuals claim to be concerned about their privacy and value it highly, their decisions are significantly inconsistent with the value they profess, as they do little or essentially nothing to protect their personal data and thus their privacy[11].

[10] Zuboff, *ob. cit.*, p. 524.

[11] Artigot Golobardes, M. "Mercados digitales, inteligencia artificial y consumidores", *El Cronista El Cronista del Estado social y democrático de Derecho*, n.º 100, 2022, pp. 130 and 131; more extensively, Barth and De Jong, "The privacy paradox -Investigating Discrepancies between expressed privacy concerns and actual online behavior -A sytematic literature review", *Telematics and Informatics*, 34(7) (2017); Norberg, P. A. and Horne D. A. "The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors", *Journal of Consumer Affairs*, 41 (1), 2007, pp. 100-126.

And, secondly, although the consolidation of an authoritarian state surveillance capitalism such as China's does not appear to be forthcoming in Europe, this does not mean that there are not already practices of corporate surveillance capitalism which, to paraphrase Zuboff again, use human experience as a free raw material for a series of hidden commercial and labour practices of extraction, prediction, and sales, presided over by a parasitic logic in which the production of goods and services and labour relations are gradually subordinated to a new global architecture of behavioural modification.

In Creemers' words, this "tendency to socially engineer and "nudge" individuals towards "better" behaviour is also part of the Silicon Valley approach, which holds that human problems can be solved once and for all through the disruptive power of technology. Human beings are reduced to a set of numbers that indicate their performance on pre-set scales, in their eating habits, for example, or in their physical exercise regime, which they are then challenged to improve. The mere fact that information exists means that companies and governments will seek to exploit it for their own purposes, whether political or commercial. In that sense, perhaps the most shocking element of the story is not the Chinese government's agenda, but how similar it is to the path technology is taking elsewhere"[12].

And to mention a specific example in Spain of the use of data that a company has been using with the aim of avoiding the requirements of a dependent employment relationship and at the same time to behaviourally "push" workers to be available as long as possible in order to obtain more orders and, in short, higher pay, it is worth recalling, even if it is a bit lengthy, what was said by the Social Chamber of the Spanish Supreme Court in its ruling of 25 September 2020 on the status of GLOVO delivery drivers as salaried workers:

> "Factual background nº7: The company has established a rating system for "glovers", classifying them into three categories: beginner, junior and senior. If a delivery driver has not accepted any service for more than three months, the company can decide to downgrade him (Clause four of the service contract). The ranking system used by GLOVO has had two different versions: the *fidelity* version, which was used until July 2017, and the *excellence* version, used from that date onwards. In both systems, the delivery driver's score is based on three factors: the final customer's assessment, the efficiency demonstrated in the com-

[12] *China's chilling plan to use social credit ratings to keep score on its citizens*, CNN, 27 October 2015, https://edition.cnn.com/2015/10/27/opinions/china-social-credit-score-creemers/index.html (as of 18 March 2024).

pletion of the most recent orders, and the performance of services during peak hours, known by the company as "diamond hours". The maximum score that can be obtained is 5 points. There is a penalty of 0.3 points each time a delivery driver is not operational in the time slot previously booked by him. If the unavailability is due to a justified cause, there is a procedure for communicating and justifying this cause, avoiding the penalising effect... The delivery drivers who have the best score have preferential access to the services or errands that are coming in...

Eighteenth legal basis: ... In practice, this system of rating each delivery driver conditions his freedom of choice of timetables because if he is not available to provide services in the time slots with the highest demand, his rating decreases and with it the possibility of being assigned more services in the future and achieving the economic profitability he is seeking, which is equivalent to losing employment and remuneration. In addition, the company penalises delivery drivers by not assigning them orders when they are not operating in the reserved slots, unless there is a justified cause that is duly communicated and accredited.

The consequence is that delivery drivers compete with each other for the most productive time slots, with economic insecurity resulting from commission-based pay with no guarantee of minimum orders, which encourages drivers to try to be available for as long as possible in order to get more orders and higher pay.

Twenty-first legal basis – Glovo is not a mere intermediary in the procurement of services between shops and delivery persons. It does not merely provide an electronic intermediary service consisting in bringing consumers (the customers) and genuine self-employed workers into contact with each other, but coordinates and organises the production service. It is a company that provides courier and messenger services by setting the price and terms of payment for the service, as well as the essential conditions for the provision of the service. And it owns the essential assets for the performance of the activity.... The company has established instructions that enable it to control the production process. Glovo has established means of control that operate on the activity and not only on the result by means of algorithmic management of the service, the valuations of the delivery drivers and constant geolocation... To provide these services, Glovo uses a computer programme that assigns the services according to the valuation of each delivery driver, which decisively conditions the theoretical freedom to choose schedules and to refuse orders. In addition, Glovo has the power to sanction its delivery drivers for a variety of different behaviours, which is a manifestation of the employer's managerial power. Through the digital platform, Glovo carries out a real-time control of the provision of the service, without the delivery person being able to carry out his task without being linked to that platform...".

Other examples in the workplace can be mentioned; Todolí Signes explains, in an extensive quote, that "work in a *call centre* is one of the most affected by this high level of monitoring. Algorithms control the number of calls attended, their duration, pauses, even the content of the call through the detection of key words, tone of voice and intonation... The company *CallMiner* announces that its software can evaluate and score -and rank workers- in terms of professionalism, courtesy and empathy in the attention shown during calls... In the same way, supermarkets can measure how fast each cashier scans the products in the shopping basket and compare them with the rest of the workers for the purposes of remuneration, assigning work shifts, dismissing those who are less fast and making cashiers compete with each other to speed up the pace of work. Computer work, whether in the office or teleworking, is another area subject to absolute control of working times and subsequent evaluation by algorithms through productivity indexes. The company Crossover offers a tool called *WorkSamart* to monitor computers. This programme counts keyboard and mouse clicks, the computer screen, emails sent and even takes a picture every ten minutes via the computer's webcam. In this way, every second of inactivity with the computer -which does not mean that the worker is not thinking or working with a notebook- is penalised...

Face-to-face jobs are not spared from such productivity checks and rankings. They exist in transport, cleaning, hospitality, etc. The best known example is Amazon's monitoring of warehouse workers by measuring the number and speed of boxes packed, the number of steps taken in a day in the warehouse, bathroom breaks, or socialising, etc. Thus, by means of smart bracelets or chips in the boots, an exhaustive count is made of the work done and, together with other variables, a productivity index is drawn up which is used to generate automatic warnings (the bracelet vibrates or a message is sent to it) or to automatically dismiss people who do not reach a minimum productivity level. According to the data, 10% of Amazon's warehouse workers in the US have been fired because of the productivity index"[13].

Finally, and to briefly approach a different area such as insurance contracts, a classic example is the use of the credit rating of the insured to set

[13] "Artificial Intelligence will not steal your job, but your salary. Retos del Derecho del Trabajo frente a la dirección algorítmica del trabajo", *El Cronista del Estado social y democrático de Derecho*, no. 100, 2022, pp. 155 and 156; more extensively, and by the same author, *Algoritmos productivos y extractivos. Cómo regular la digitalización para mejorar el empleo e incentivar la innovación*, Aranzadi, 2023.

the premium in motor insurance, which, as María Luisa Muñoz Paredes recalls, gave rise to a rejection movement in the United States, following the finding by the Consumer Reports Association in 2015 that this factor was taken into account more than other more influential factors in risk, such as the driving record of the insured[14]. In this regard, Recital 37 of the AIA recalls that AI systems intended to be used for risk assessment and pricing in relation to individuals for health and life insurance can also have a significant impact on people's livelihoods and, if not properly designed, developed and used, can violate their fundamental rights and lead to serious consequences for people's lives and health, including financial exclusion and discrimination.

With the provisions contained in the Regulation, some of these tools, as mentioned above, will be considered "high risk" systems if the data used come from the context in which the results of the evaluations are applied, and may be prohibited if they come from different contexts and generate discrimination.

IV. The prohibition of certain systems that evaluate or classify natural persons

Article 5.1(c) of the Regulation has had the following course from the Commission's proposal of 21 April 2021 to the final wording, before the Common Position ("general approach") of the European Council on the AIA of 6 December 2022 and the amendments formulated by the European Parliament on 14 June 2023.

The following AI practices are prohibited

[14] " Big Data, AI y seguro: riesgos de inasegurabilidad y discriminación entre asegurados", *El Cronista del Estado social y democrático de Derecho*, n.º 100, 2022, p. 122; more extensively, and by the same author, ""Big Data" y contrato de seguro: los datos generados por los asegurados y su utilización por los aseguradores", in Huergo Lora, A. H (dir.): *La regulación de los algoritmos*, Aranzadi, Cizur Menor, 2020, pp. 129-162; "El "Big Data" y la transformación del contrato de seguro", in Veiga, A. B. *Dimensiones y desafíos del seguro de responsabilidad civil*, Cizur Menor (Aranzadi), 2021, pp. 1017-1051; on the use in insurance contracts of what Caty O'neil calls "weapons of mathematical destruction" see her book of the same title, Capitán Swing, Madrid, 2017, pp. 199 ff.

Commission	*European Council*	*Parliament*	*Regulation*
The placing on the market, putting into service or use of AI systems by or on behalf of public authorities for the purpose of assessing or classifying the reliability of natural persons over a given period of time on the basis of their social conduct or known or predicted personal or personality characteristics, in such a way that the resulting social ranking results in one or more of the following situations: (i) prejudicial or unfavourable treatment of particular individuals or entire groups in social contexts which are unrelated to the contexts in which the data were originally generated or collected; (ii) prejudicial or unfavourable treatment of certain individuals or entire groups which is unjustified or disproportionate to their social behaviour or the gravity of the latter.	The placing on the market, putting into service or use of AI systems for the purpose of assessing or ranking natural persons over a given period of time on the basis of their social behaviour or known or predicted personal or personality characteristics, in such a way that the resulting citizen score results in one or more of the following situations: (i) prejudicial or unfavourable treatment of particular natural persons or groups of natural persons in social contexts which are unrelated to the contexts in which the data were originally generated or collected; (ii) prejudicial or unfavourable treatment of certain natural persons or groups of natural persons which is unjustified or disproportionate to their social behaviour or the gravity of the latter.	The placing on the market, putting into service or use of AI systems for the purpose of assessing or ranking natural persons or groups of natural persons for social rating over a given period of time on the basis of their social behaviour or known, inferred or predicted personal or personality characteristics, in such a way that the resulting citizen score results in one or more of the following situations: (i) prejudicial or unfavourable treatment of particular individuals or entire groups in social contexts which are unrelated to the contexts in which the data were originally generated or collected; (ii) prejudicial or unfavourable treatment of certain natural persons or groups of natural persons which is unjustified or disproportionate to their social behaviour or to the gravity of the latter.	The placing on the market, putting into service or use of AI systems for the purpose of assessing or ranking natural persons or groups of persons over a given period of time on the basis of their social behaviour or known, inferred or predicted personal or personality characteristics, in such a way that the resulting citizen score results in one or more of the following situations: (i) prejudicial or unfavourable treatment of particular individuals or entire groups of individuals in social contexts which are unrelated to the contexts in which the data were originally generated or collected; (ii) prejudicial or unfavourable treatment of certain natural persons or groups of persons which is unjustified or disproportionate to their social behaviour or to the gravity of the latter.

Table prepared by the authors.

Although this is not one of the provisions that has undergone most changes between the Commission's proposal and the amendments adopted by Parliament, it is worth highlighting those that have been made and, first of all, one of the most important is the one relating to the person prohibited from introducing these systems: whereas the Commission's proposal mentioned "public authorities" or anyone acting "on their behalf", the Council's common position, as well as Parliament's amendment and the final wording resulting from the interinstitutional agreement remove this specification and the prohibition will affect both public authorities and private individuals, whether physical or legal, including, therefore, companies.

This modification seems very positive because the risks to be combated can come from both public and private parties and, as we have already seen, we find examples of the use of scoring systems by very important companies.

Secondly, the Commission's proposal referred to the assessment or classification of the 'trustworthiness' of natural persons, whereas the Council's common position, Parliament's amendment and the final text refer to 'assessing or classifying natural persons or groups of persons', i.e., the analysis is not limited to the "trustworthiness" of a person but extends to the person as such and, moreover, Parliament's amendment includes persons "or groups of persons" (e.g., consumer groups, workers, insured persons, etc.).

Thirdly, the Commission and Council texts, although not identical -the former refers to 'social conduct or personal haracteristics or personality traits' and the latter to 'social behaviour or personal characteristics or personality traits'- refer to 'known or predicted' characteristics, whereas the Parliament's amendment and the final wording of the Regulation also include 'inferred' characteristics, which is relevant because inferences are conclusions drawn from data processing and this is one of the properties of AI systems: the ability to extract new information from existing data.

Fourthly, while the Commission's proposal speaks of "social ranking", the Council and the Parliament use the term "citizen score", which will finally be included in the "Regulation", although it does not seem that the idea to which they refer is different: the ranking of people on the basis of known, predicted or inferred data.

The fifth issue to comment on is the generation of one or more of the situations described below that would justify the prohibition, the first of which is that it results in detrimental or unfavourable treatment of specific individuals or entire groups in social contexts unrelated to those in which the data were originally generated or collected. The score resulting from the processing of the data is considered to result in discrimination or, in the words of the texts under consideration, "detrimental or unfavourable treatment".

In this respect, and as we have seen at the beginning, the final wording of Recital 31 explains that "AI systems providing social scoring of natural persons by public or private actors may lead to discriminatory outcomes and the exclusion of certain groups. They may violate the right to dignity and non-discrimination and the values of equality and justice."

A significant qualification, to which we have already referred to above, is that the data generating such unfavourable treatment must have been obtained in contexts other than the one in which they would cause the detriment, but nothing would prevent their use in the context of origin; In this respect, it seems that data obtained in the context of an employment relationship could be used to carry out a scoring of those who work in that company or data obtained in a contractual relationship for the provision of services (for example, electricity supply) to establish a hierarchy of different prices to customers in different situations because one thing is the difference in prices and another discrimination; in this line, Law 3/1991, of 10 January, on Unfair Competition, in article 16.1 establishes that "discriminatory treatment of the consumer in terms of prices and other conditions of sale shall be considered unfair, unless there is a justified cause", i.e., different treatment for which there is justification would not be unfair, nor would the mere difference in prices[15].

However, the absence of discrimination or detrimental treatment contrary to the prohibition of Article 5.1(c) does not exclude that the data used are used without the knowledge or even the consent of the person concerned, which may place him in a position of particular vulnerability in digital markets. For this reason, "it is necessary to create mechanisms to prevent such vulnerability from materialising in the form of an expropriation of the contractual surplus that the consumer expected to obtain from the transaction and that only purely contractual instruments will not be able to recover"[16].

On the other hand, and as also noted above, the fact that the system in question is not subject to prohibition does not exclude that it can be qualified as "high risk" in the terms already seen.

Finally, and as has already been pointed out, what would be unacceptable is the use of someone's data to carry out evaluations or classifications in a context other than the one in which they were generated or obtained and

[15] See in this regard Muñoz Paredes, M. L. "Big Data, AI y seguro: riesgos de inasegurabilidad y discriminación entre asegurados", *El Cronista del Estado social y democrático de Derecho*..., p. 123.

[16] Golobardes, A. "Mercados digitales, inteligencia artificial y consumidores", *El Cronista El Cronista del Estado social y democrático de Derecho*...p. 135.

which would entail prejudice or unfavourable treatment[17]; thus, for example, a person's higher or lower credit rating should not be a conditioning factor for promotion within a company[18].

The second scenario that would justify the prohibition of an AI system is if it leads to "detrimental or unfavourable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behaviour or the seriousness of their behaviour". What is taken into account here is the way in which a natural person interacts with and influences other natural persons or society, resulting in unfavourable treatment that is either unjustified or the consequences are disproportionate to its severity; for example, that political opinions or ideological, religious, social, or cultural manifestations expressed on a social network generally imply a cause for exclusion from employment or expulsion from an educational establishment or that ratings of a worker's friendliness by customers are sufficient cause for dismissal or an unreasonable financial penalty.

[17] Obviously, comments or behaviour that are in breach of contractual good faith or offensive to the employer can have repercussions on the employment relationship (54.2 (c) and (d) of the Workers' Statute).

[18] Cathy O'NEIL provides numerous examples of the perverse results of the use of, among others, credit rating criteria in the labour and consumer spheres in *Weapons of Mathematical Destruction. How big data increases inequality and threatens democracy*... pp. 181 ff.

THE CONTENT OF THE SO-CALLED "SUBLIMINAL TECHNIQUES" AND THE VULNERABILITIES OF SPECIFIC GROUPS OF PEOPLE IN THE ARTIFICIAL INTELLIGENCE ACT

Luis Miguel González de la Garza[1]
Reader in Constitutional Law UNED

I. Introduction

In this chapter, we will analyse paragraphs (a) and (b) of article five of the AI Act, whose most notable notes are perhaps the concepts of *"subliminal techniques" and "vulnerabilities of a natural person or a specific group of persons."* Before doing so, however, it is worth noting where is this article systematically located in the framework of the standard. The use of AI, with its specific characteristics such as: opacity, complexity, data dependency, autonomous behaviour, can negatively and seriously affect a number of fundamental rights and the security of users. To address these concerns, the AI Act follows a sensible risk-based approach whereby legal intervention is tailored to the specific level of risk. To that end, the AI Act distinguishes between AI systems that present (i) unacceptable risk (ii) high risk (iii) limited risk and (iv) low or minimal risk. AI applications would only be regulated to the strictly necessary extent to address specific levels of risk. Chapter II (Article 5) of the AI law explicitly prohibits harmful AI practices that are considered a clear threat to people's safety, livelihoods and rights, due to the "unacceptable risk" posed by their use. Consequently, it would be prohibited to market, provide services or use such practices in the EU.

It follows from the above that we are in the presence of "unacceptable risk" techniques, i.e., the most restrictive techniques provided for by the rule. But, apart from what we will see below, what are -synthetically- these techniques that constitute an unacceptable risk? They are essentially techniques or forms of mental manipulation aimed at substantially or significantly altering the behaviour of an individual or a group of persons by altering their ability to form preferences by means of behavioural strategies that are known or may be developed in the future and in which AI systems are suitable for their application. The article under consideration does not describe them, but out-

[1] This work is the result of the research project "Educating in values, building citizenship", Ministry of Science and Innovation. State Research Agency. Knowledge Generation Projects 2021. Referencia: PID2021-127680OB-I00.

lines the essential principles for their identification and realisation, since they can be deployed in different ways and, above all, according to very different qualitatively technologies.

In our opinion, there is no doubt about the timeliness and appropriateness of the need for this regulatory provision since, as we shall see, two groups of technologies operate synergistically in the field of risks to people, the first being AI with its immense capacity to process precise *quantitative* data on individuals, groups, or collectives characterised by common traits, for example psychological ones, and, on the other hand, the development of neurotechnologies is not intended for the medical treatment of patients but rather for uses that appear to be recreational or therapies not covered by medical regulations. However, under lax regulations – what happens particularly in the United States, as Farahany warns[2] – they can obtain both mental data and modify behaviors through the generation of electromagnetic fields in response to the processing of mental data processed through AI. Separate but related to the above are medical technologies based on direct access to the brain through bio-implants, in which an essential phase of information processing will be carried out by AI. These technologies have a very significant disruptive capacity as they have the characteristic of permanence – as they are fixed implantation systems – and not a one-off feature like technologies based on radio frequencies.

II. Developments in processing and content

We will now consider the development of paragraphs (a) and (b) of Article 5, which are the subject of our commentary, therefore, we will focus on the AI Act Proposal, amending certain Union legislation of 19 October 2022, document from the Presidency to the Delegations. In this text we pay attention to recital 16 which states:

> *"AI-enabled manipulative techniques can be used to persuade persons to engage in unwanted behaviours, or to deceive them by nudging them into decisions in a way that subverts and impairs their autonomy, decision-making and free choices. The placing on the market, putting into service or use of certain AI systems intended to distort materially distorting human behaviour, whereby physical or psychological harms are likely to occur, are particularly dangerous and should therefore be forbidden. Such AI systems deploy subliminal components such as audio, image, video stimuli individuals that persons cannot perceive as those stimuli are beyond human*

[2] Farahany, Nita A, *The Battle for your Brain. Defending the right to think freely in the age of neurotechnology*, St. Martin's Press, 2023, New York, pp. 29-35.

> *perception or other subliminal -making or free choices in ways that people are not consciously aware of, or even if aware not able to control or resist, for example in cases of machine-brain interfaces or virtual reality.*
>
> *In addition, AI systems may also otherwise exploit the vulnerabilities of a specific group of persons due to their age, disability within the meaning of Directive (EU) 2019/882, or a specific social or economic situation that is likely to make those persons more vulnerable to exploitation such as persons living in extreme poverty*[3] "

As we can see, a guiding idea is based on the fact that the vector of application of these technologies is audiovisual systems, i.e., screen technology such as computer screens, smartphones in all their possible configurations, including Muse-2 type headsets among others, or *augmented reality* glasses and *virtual reality* glasses, which are different from the former, such as the Apple Vision Pro or Meta Quest 3, among other technologies. There is also a brief reference to brain-computer interface, better known as BCIs, to which the EU legislator will devote much more attention in future amendments to the regulation, particularly in the recitals, but not in the legal text.

The focus is on children and specific groups of people who, because of their age or varying abilities, should be particularly protected as they are more vulnerable to the potential use of such technologies, as defined in Article 3(1)[4] of Directive (EU) 2019/882.

With regard to the regulatory content, the Commission's wording of 19 October 2022 is expressed in the following terms:

*Article 5 […] 1.*The following Artificial Intelligence practices shall be prohibited:

(a) the placing on the market, putting into service or use of an AI system that deploys with the objective to or the effect of in order to materially distorting causes or is reasonably likely to cause that person or another person physical or psychological harm;

(b) the placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability or a specific social or economic situation, with the objective to or the effect of in order to materially distorting the behaviour of a person pertaining to that group in a manner that causes or is reasonably likely to cause that person or another person physical or psychological harm;

[3] Italics are ours in the texts correspond to bold.

[4] "persons with disabilities" means persons who have long-term physical, mental, intellectual or sensory impairments which in interaction with various barriers may hinder their full and effective participation in society on an equal basis with others.

On 6 December 2022 the General Secretariat of the Council adopted a new text for the delegations, which contains some interesting changes. Starting with the recitals, it is considerably more detailed than that of 19 October 2022 and is numbered identically:

> "AI-enabled manipulation techniques can be used to persuade people to adopt unwanted behaviours or to trick them into making decisions in a way that undermines and damages their autonomy, decision-making and ability to make free choices. The introduction on the market, putting into service or use of certain AI systems that substantially alter human behaviour, making physical or psychological harm likely, are particularly dangerous and should therefore be prohibited. Such AI systems use subliminal components, such as sounds, images or video stimuli, which people cannot perceive, as such stimuli transcend human perception, or other subliminal techniques that undermine or impair people's autonomy, decision-making or ability to make free choices in ways that people are not really aware of, and even if they are aware of them, cannot control or resist them, for example in the fields of brain-machine interfaces or virtual reality. In addition, AI systems may also otherwise exploit the vulnerabilities of a specific group of people, stemming from their age, their disability within the meaning of Directive (EU) 2019/882 or a specific social or economic situation that may make them more vulnerable to exploitation, such as people living in extreme poverty or ethnic or religious minorities. Such AI systems may be placed on the market, put into service or used with the purpose or effect of substantially altering the behaviour of a person and in a way that causes or is reasonably likely to cause physical or psychological harm to that person or to another person or group of persons, in particular harm that may accumulate over time. The intent to distort behaviour cannot be assumed if the disturbance is the result of factors external to the AI system that are beyond the control of the provider or user, i.e., factors that the provider or user of the AI system cannot reasonably foresee or mitigate. In any case, the provider or user need not intend to cause the physical or psychological harm, provided that such harm arises from manipulative or exploitative AI-enabled practices. The prohibitions of such AI practices complement the provisions of Directive 2005/29/EC, in particular the prohibition, in all circumstances, of unfair commercial practices that cause economic or financial harm to consumers, whether established through AI systems or otherwise. The prohibition of manipulative and exploitative practices contained in this Regulation should not affect lawful practices in the context of medical treatment, for example psychological treatment of mental illness or physical rehabilitation, where such practices are carried out in accordance with applicable medical standards and legislation. Furthermore, common and legitimate commercial practices, in conformity with the applicable law, should not be considered as harmful AI manipulation practices per se".

As we can see, apart from specific twists in the wording of the wording of 19 October 2022, it is added that the prohibition of such AI practices will complement the provisions of Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market, which amends Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No. 2006/2004 of the European Parliament and of the Council, for example with regard to the prohibitions set out in Article 5.3, which states:"*Commercial practices that may substantially distort, in a manner that the trader can reasonably foresee, the economic behavior solely of a clearly identifiable group of consumers particularly vulnerable to such practices or to the product to which they refer, due to the latter suffering from a physical ailment or a mental disorder, or to their age or their credulity, should be assessed from the perspective of the average member of that group. This will be understood without prejudice to the usual and legitimate advertising practice of making exaggerated claims or claims for which a literal interpretation is not intended*".

Excluded from the aforementioned prohibited practices are those legal practices in the context of medical treatment, citing as an example the psychological treatment of a mental illness or physical rehabilitation, when such practices are carried out in accordance with the rules, i.e., the *lex artis* of mental health professionals. Let us think of treatments in which AI could be used with patients who need to handle information models appropriate to the pathologies they suffer from. In these cases, the professional associations may have a responsibility in the knowledge, supervision and adequacy of these practices in their use by professional members, a responsibility that may be shared with university training centres. Recital 16 also points out that common and legitimate commercial practices, in accordance with the applicable law, should not be considered as harmful AI manipulation practices in themselves. In Spain, this prohibition is found in Article 3(c) of Law 34/1988 of 11 November 1988 on General Advertising and is defined in Article 4, which states: "For the purposes of this law, subliminal advertising is that which, by means of techniques for the production of stimuli of intensities bordering on the thresholds of the senses or similar, can act on the target public without being consciously perceived". It is also in the fourth paragraph[5] of Article 122 "Absolute prohibitions of certain audiovisual communications" of Law 13/2022, of 7 July, General Audiovisual Communication. The concepts

[5] 4. Subliminal audiovisual commercial communication which, by means of techniques for the production of stimuli of intensities bordering on the threshold of the senses or similar, may act on the target audience without being consciously perceived, shall be prohibited.

of *subliminal advertising* must be carefully distinguished from other borderline concepts in order not to confuse them with those of *surreptitious advertising* and *suggestive advertising*, which are referred to in the final part of recital sixteen and which are legitimate neuromarketing strategies, Diotto[6]. On these distinctions, see Tato Plaza.[7]

With regard to the articles of the drafting of 6 December 2022, the wording undergoes non-substantial variations, modifying 'commercialisation' by 'placing on the market' in such a way that the frontier of access to subliminal technology is advanced.

Finally, we will look at the text adopted by the European Parliament and the amendments adopted on 14 June 2023 on the AI Act Proposal.

Let us first look at the changes made to Recital 16 by amendment 38. One of the most relevant elements of the Parliament's wording is the introduction of BCI-connected brain prostheses no longer as in the previous wording; here it is now expressly stated that: "This *limitation* should be understood to *cover neurotechnologies assisted by AI systems* that are used to *monitor, use, or influence neural data collected through brain-computer interfaces, to the extent that they substantially alter the behaviour of a natural person in a way that causes or is likely to cause significant harm to that person or to another person.* In other words, the The The European Parliament clearly takes into consideration the interrelation between neurotechnologies and Artificial Intelligence, which seems important to us because one of the major areas in which AI can substantially alter people's behaviour will be determined by neurotechnologies and the fact that they operate through AI systems is and will be the norm due to the immense complexity of data that needs to be processed to collect and process – electrically translate – the output signals as well as the input signals to the brain. That said, there is a lack of ambition to have connected this type of treatment with the so-called neuro-rights that we will see later on.

It is important to distinguish different areas addressed by the wording we are considering. Subliminal techniques could be employed through display systems of various types of devices, as we considered above; however, manipulation techniques through data processing are a very different aspect as they would involve operating on the electrochemical inputs, which would be translations of the signals manipulated by the AI, so that the consciousness would not even have the opportunity to detect such manipulations, and there-

[6] Diotto, M, "*Neuromarketin. Las herramientas técnicas de una estrategia de marketing eficaz para creativos y especialistas en marketing*", Hoepli Ediciones, Madrid, 2022, pp. 131-157.

[7] Tato Plaza, A, in J A, García-Cruces, *Tratado de Derecho de la Competencia y de la Publicidad*, Tomo II, Tirant lo Blanch, Valencia, 2014, pp. 1964-1967.

fore, in our opinion, this is a qualitatively different type of manipulation from subliminal techniques, which we will discuss in more detail later on.

Regarding the wording of the text of 14 June 2023, it is important to consider amendment 215 corresponding to paragraph 1(a) and amendment 216 paragraph 1(b) partially redrafting Article 5. The text of these amendments clearly identifies subliminal techniques utilising AI. It also refers to *techniques which are deliberately manipulative* or *deceptive* with the purpose or effect of substantially altering the behaviour of a person or group of persons by appreciably impairing their ability to make an informed decision and thereby causing the person to take a decision that he or she would not otherwise have taken. The latter techniques employ AI but are not necessarily subliminal and could include neurotechnology-based techniques through BCI and AI systems and other less invasive techniques such as *priming* or all those that exploit known *cognitive biases* Garrigues and de la Garza[8] that aim to exploit automatic thought forms characteristic of *unconscious* information processing. The last paragraph points out the prohibition of AI systems making use of subliminal techniques, which shall not apply to AI systems intended to be used for therapeutic purposes authorised on the basis of a specific informed consent of the persons exposed to them or, where appropriate, of their legal guardian, i.e., in cases of therapeutic forms using AI in the framework of mental health as we saw above.

Letter b) generally retains the wording of 6 December 2022, which in turn is derived from 19 October 2022, although it adds: including known or predicted characteristics of personality traits or the social or economic situation of that person or group, age, and physical or mental capacity. In this sense, we can think about the limitation of processing by AI-based systems that, through Big Data, process personality information but also economic information or both. This information is that which has already been used in the fields of virtual cognitive electoral propaganda through microtargeting.

Finally, on 14 March 2024, the AI Act adopted by the Parliament was published.[9] In this latest amendment, Recital 16 is changed to Recital 29, and there are no substantial modifications with respect to the previous wording; the concern for emerging neurotechnologies in relation to AI and their capacity

[8] Garrigues Walker, A, González de la Garza, L. M, *El derecho a no ser engañados. Y cómo nos engañan y nos autoengañamos*, Thomson Reuters Aranzadi, Navarra, 2020, p. 87.

[9] Artificial Intelligence Regulation. European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)).

to produce substantial modifications in people's behaviour is maintained. We believe it would have been prudent not only to consider the effects of BCIs (*brain-computer interfaces*) but also to contemplate the influence through induced electromagnetic fields that could be the current equivalent, in the 21st century, of the subliminal 20th century techniques. Twentieth-century subliminal brain access techniques occur through the visual and auditory systems; those of the 21st century will add BCI in the form of neuroimplants and headbands by reading and modifying behaviour through electromagnetic fields processed by AI, as studied in detail by Garrigues and de la Garza[10]. Computational propaganda is a reality, as pointed out by Wooley and Howard, among others, although the novelty of the type of propaganda we are considering is not that it is passive propaganda, but rather propaganda that we could call active and intelligent, because it takes advantage of the *characterological biases* of voters to design a campaign of very high granularity and precision, tailored precisely to the voter and his or her emotional and political preferences. If, for example, it is a voter who has abstained from voting in previous elections, it is possible to offer them arguments based on their emotional preferences to vote. We can think of voters who exhibit traits that can be exploitable by automated propaganda agents, voters who do not have a clear preference and whom this type of propaganda can "follow" in such a way that through "*microtargeting*" or micro-segmentation, it seeks out the voter to offer them active propaganda of their liking, capable of learning from the interaction with the voter based on their personality and readapting and *refining itself* according to the voter's responses in a virtual dialogue of propaganda accompaniment directed by *predictive* AI that was non-existent prior to the advent of these technologies.

It is called microtargeting because it aims to group voters into very small segments or *clusters*[11] synchronised with the 20 models of personality types or *psychometric profiles* already elaborated and targeted by this type of electoral propaganda. This ensures that the personalised information reaches its intended electoral audience. It is usual to observe in any Internet browsing that after visiting a virtual shop, information about the product or service that we have visited previously, in previous hours, days or weeks, appears on our computers or mobile phones, advertising follows the browser on certain

[10] Garrigues Walker, A, L M, González de la Garza, *Qué son los neuroderechos y cuál es su importancia para la evolución de la naturaleza humana*, Aranzadi, Navarra, 2024.

[11] Socio-demographic, such as contested constituencies or specific constituencies, where few votes can produce the allocation of a seat and where a cognitive propaganda activity can justify an extra campaign effort to lead voters who are unsure whether or not they will exercise their vote to be motivated towards a particular electoral tendency.

Web pages thanks to the use of cookies[12] previously accepted and installed on the users' equipment in which this contextual advertising "that searches for us and accompanies us" appears. This tracking would be the equivalent of electoral microtargeting in its commercial dimension. But unlike commercial microtargeting, the election campaigner talks to and learns from the voter. He will then try to persuade the voter with logical and emotional arguments by trying to mimic the voter's personal, social, and emotional interests and offering him different versions of the propaganda campaign that are tailored to his psychological profile. Experiments on the manipulation and mass contagion of emotions in social networks, such as the one on Facebook in 2012 involving 700,000 subjects, as studied by Kramer, Guillory, and Hancock (8788-9790:2014)[13], convincingly demonstrate the great effectiveness of what can be achieved in the field of transforming motivations and preferences by means of induced emotional contagion.

In Donald Trump's 2016 election campaign, Cambridge Analytica (now Emerdata), as Cadwalladr[14] points out, was employing between forty and fifty thousand variants of different informative election pitches whose response was measured in real time by the recipients, readapting to their responses in an evolving way. The granularity of the actions of these messages is structured by geographical areas of up to a radius of 5 miles in which psychographic profiles are grouped,[15] which are evaluated by the Cambridge Analytica algorithm whose origin is at the University of Cambridge[16] and which uses OCEAN[17] to analyse the personality types of the voters it subsequently seeks to influence. In addition, the variants of the propaganda messages currently used cannot be known to other voters, as they are based, for example,

[12] López Jiménez, D. Las cookies como instrumento para la monitorización del usuario en la red: La publicidad personalizada, *Ciencias Económicas* 29, n.º 2, 2011.Socio-demographic,

[13] Kramer, Adam D.I, J E. Guillory and J T. Hancock, Experimental evidence of massive-scale emotional contagion through social networks, *PNAS*, Vol. 111, No. 24, 17 June 2014.

[14] Cadwalladr, C, Google, "democracy and the truth about internet search" Internet, The Observer, 4 December 2016. https://www.theguardian.com/technology/2016/dec/04/google-democracy-truth-internet-search-facebook (20 March 2024).

[15] Psychographic segmentation is a tool that makes it possible to delve deeper into reference groups to find their voting motivations.

[16] The reader can experience a basic psychographic analysis of their social networks with this algorithm at: https://www.psychometrics.cam.ac.uk/productsservices/apply-magic-sauce.

[17] According to Goldberg, the five major personality traits, also called core factors, are named as follows: Factor O (openness to new experiences), Factor C (responsibility), Factor E (extroversion), Factor A (agreeableness) and Factor N (neuroticism or emotional instability), thus forming the acronym "OCEAN".

on Facebook, on invisible[18] *or dark posts*, which were and are initially a tool for personalised advertising, but which can also be used in personalised cognitive election campaigns and which are difficult for a future electoral authority to monitor.

We adhere to the conclusions that Wolley and Howard[19] point out. In this sense, computational propaganda is one of the most powerful tools against democracy since it makes possible a genuine and new form of "*social engineering*" capable of completely breaking the patterns of public opinion and its manipulation as studied by Bond, Fariss, and collaborators[20]. Indeed, the electoral cognitive propaganda systems seem to work in parallel to powerful and profound distortions of public opinion that are being originated by very diverse interest groups of national and international scope capable of modifying, for example – through computer farms – the agenda of public opinion on issues of political interest through the manipulation of trends based on the generation of hashtags to achieve positioning as Trending Topics, as Nimmo[21] points out. However, these trends are created artificially and intentionally by means of AI, both by the aforementioned computer farms and by automated *bots*[22] as Ferrara[23] or other technological vectors of generation and dissemination at the service of their creators. The phenomenon has been studied by Bradshaw and Howard[24] in the international context, and a very worrying body of evidence has been found, since the main task of these platforms, which was originally to shape public opinion through the use of "dynamic narratives" to combat the propaganda disseminated on the networks by terrorist organisations, has now shifted to other completely

[18] https://www.facebook.com/business/a/online-sales/unpublished-page-posts (20 August 2023).

[19] Woolley, Samuel C, and Philip N. Howard, *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*, Oxford Studies, 2018.

[20] Bond, Robert M, Christopher J. Fariss, Jason J. Jones, AdamD.I.Kramer, Cameron Marlow, Jaime E. Settle& James H. Fowler, A 61-million-person experiment in social influence and political mobilization, Nature, Vol 489, 13 September 2012.

[21] Nimmo, B, *Measuring Traffic Manipulation on Twitter*, Computational Propaganda Research Project, Oxford University, 2019.

[22] For a taxonomy of the various types of Bots suitable for social engineering, see: Ferrara, E, Onur Varol, C Davis, F Menczer and A Flamini, The Rise of Social Bots, *Communications of the ACM*, July 2016, Vol. 59, No. 7.

[23] Ferrara, E. et al., The Rise of Social Bots, Communications of the ACM, July 2016, Vol. 59, no. 7.

[24] Bradshaw, Samantha and Philip N. Howard, Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation, Working paper no. 2017.12, Computational Propaganda Research Project, Oxford University, UK.

different activities, such as those of a political nature, as the effectiveness or efficiency of these techniques has been demonstrated for purposes other than those for which they were originally designed.

This refers to the elaboration of false information, which is then disseminated or vectored on social networks by groups or individuals. Guess, Nagler and Tucker[25] have recently studied which social groups – by age – are the most characteristic agents of dissemination on networks such as Facebook, concluding that a small percentage of Americans, less than 8.5 percent, shared links to "fake news" sites during the 2016 election campaign, but this behaviour was disproportionately common among people over the age of 65 regardless of ideological or political affiliation, with younger people playing a much smaller role. These previously unregulated behaviours can be efficiently prohibited with the wording we have considered.

III. What are subliminal techniques?

Psychologists have long known that weak, degraded or short-lived stimuli are often not consciously perceived, but can nevertheless lead to responsive behaviour, as Edelman and Tononi point out[26] More than forty years ago Vance Packard in his bestselling book *The Hidden Persuaders* made this subliminal perception popular with his famous message "Drink Coca-Cola[27] " which was shown very briefly during a film screening with the intention of arousing the thirst of viewers without them consciously recognising the message. For many years the weak scientific evidence in support of subliminal perception was the subject of much scepticism, but subsequent studies have established the phenomenon through controlled experiments. In the laboratory, subliminal perception – now often referred to as *unconscious perception* – is usually demonstrated by the presentation of stimuli that are too weak, short or noisy to be consciously perceived, but are sufficient to rush or bias the subject's ability to perform a lexical decision task or equivalent tests. Edelman and Tononi[28] remind us, for example, that if the word *walk* is shown for a very

[25] Guess A, J Nagler and J Tucker, Less than you think: Prevalence and predictor of fake news dissemination on Facebook, Sci. Adv. 2019; 5: eaau4586 9 January 2019. Available from: https://advances.sciencemag.org/content/advances/5/1/eaau4586.full.pdf%20 (8 May 2022).

[26] Edelman Gerald, M and G Tononi , *El universo de la conciencia*, Crítica, Barcelona, 2005 , p. 88.

[27] There are authors who doubt that the Coca-Cola experiment really existed.

[28] Op Cit, p. 89.

short time, the person will deny having seen anything; if asked later for a word that matches *bank*, the person is more likely to answer *seat* than *money*. It seems clear, then, that subliminal stimuli produce enough neural activation to trigger an appropriate behavioural response. However, there is something in the neural activation – the authors remind us – produced by such stimuli that is inadequate or insufficient for a conscious experience to emerge. ¿What are we lacking?

A series of experiments begun some 30 years ago by Benjamin Libet shed some light on this question. In one of them, Libet sent electrical impulses at 72 pulses per second through electrodes chronically implanted in the patient's thalamus for therapeutic pain control. Stimulation of certain parts of the thalamus is known to activate neural pathways that deal with tactile stimuli and produce a readily identifiable sensation. The most surprising discovery was that such weak stimuli required a remarkable amount of time of appropriate brain activity, about 500 msec (half a second) before producing a conscious sensory experience.

Libet showed that the conscious intention to act only appears after a delay of about 350 msec from the beginning of the specific brain activity preceding a voluntary act. He concluded that the brain initiation of a spontaneous and free voluntary act can begin unconsciously, i.e., before the subject is consciously aware that his or her intention to act has already been initiated in the brain Edelman and Tononi[29] .

Carl Jung wrote that "there are certain events of which we are not consciously aware, which remain, so to speak, below the threshold of consciousness. They occur but are absorbed subliminally". For Mlodinow[30] our subliminal brain is invisible to us, but it influences our conscious experience of the world in the most fundamental of ways: in how we see ourselves and how we see others, in the meanings we attach to the everyday events of our lives, in our ability to make snap judgements and decisions that sometimes mean the difference between life and death, and in the actions we initiate as a result of all these instinctive experiences.

Zimmerman[31] estimates that the human sensory system sends about eleven million bits of information to the brain every second, but our conscious mind cannot process such an enormous amount of information, which has

[29] Op Cit, pp. 90-91.

[30] Mlodinow, L, *Subliminal. How your unconscious governs your behaviour*, Crítica, Barcelona, 2018, p. 11.

[31] Zimmerman M, The Nervous System in the Contex of information theory, in R.F. Schmidt and G. Thews, eds, *Human Phsycholoy,* Springer, Berlin, 1989, pp. 166-176.

been estimated to be between sixteen and fifty bits per second, so the conscious mind cannot handle the immense amount of information processed by the *unconscious* system. Evolution, says Mlodinow[32] has endowed us with an unconscious mind because the unconscious is what allows us to survive in a world that requires us to process an immense amount of information. Sensory perception, recalling memories, activities, everyday decisions and judgements all seem effortless, but only because the effort they require is mostly made in parts of the brain that function outside our consciousness. This automation as described by Edelman and Tononi[33] so pervasive in our adult lives suggests that conscious control is only exercised at critical moments, when a specific decision or plan needs to be made. In between, countless unconscious routines are executed that allow consciousness to float free from the shackles of all these details and to engage in the task of teasing out meaning and making plans for action within the overall scheme of things. It seems that in both action and perception, only the highest levels of control and analysis are available to the consciousness: everything else is executed automatically. This feature has led many to believe that we are aware of the *results* of the brain's *operations*, but not of the operations themselves.

But in addition to the above and as Metzinger points out[34] according to several scientific studies, our minds are wandering between 30 and 50% of our conscious waking phases. If we take into consideration all empirical findings concerning *mind wandering*, we reach a surprising result of a philosophical relevance that cannot be overstated: mental autonomy is the exception; loss of control is the rule. A number of empirical studies show that the areas of the brain involved in mind wandering – i.e., conscious but referentless mental states – overlap on a large scale with the well-known *default-mode* network. The *default mode* network is usually activated during periods of rest and, as a result, attention is directed inwards. This is what happens for example, when we daydream, when unexpected memories occur, or when we think about ourselves or the future. The moment a concrete task arises, this area of the brain is deactivated and we immediately concentrate on the problem to be solved. Metzinger's hypothesis is that the default mode serves primarily to keep the autobiographical self-model stable and in good shape, like an automatic maintenance programme, generating renewed stories to make us believe that we are one and the same person over time, i.e., creating psychological continuity.

[32] Op Cit, Mlodinow, L, *Subliminal. How your unconscious governs your behaviour*, p. 45.

[33] Op Cit, Edelman Gerald, M and G Tononi, *The Universe of Consciousness,* p. 75.

[34] Metzinger, T, *The tunnel of the self. Ciencia de la mente y mito del sujeto.* Enclave, Madrid, 2018, pp. 170-171.

Bernard Baars elaborated in 1988 the *global work area theory* as outlined by Kandel[35] According to this theory, consciousness involves the diffusion of previously unconscious (preconscious) information through the cerebral cortex. Baars suggest that the global work area comprises a system of neural circuits extending from the brainstem to the thalamus and from the thalamus to the cortex. French cognitive neuroscientist Stanislas Dehaene extrapolated Baars' psychological model to the biological model. Dehaene discovered that what we perceive as a conscious state is the result of a set of neural circuits that select data, amplify it, and distribute it throughout the cerebral cortex. Baars' theory and Dehaene's findings show that we have two distinct ways of thinking about things: one is *unconscious* and involves perception; the other is conscious and involves the diffusion of perceived information. What happens in the brain when we see a word subliminally, below the level of consciousness? First of all, the visual cortex becomes very active. This is unconscious neural activity: the word we have seen reaches the primary visual processing centre of the cerebral cortex, but after 200 or 300 milliseconds, it slowly disappears without reaching the higher centres of the cortex. When a percept becomes conscious, another scenario occurs. Conscious perception also starts with signs of activity in the visual cortex, but that activity, instead of diminishing, intensifies. After about 300 milliseconds, it is very intense; it is like a *tsunami*, not a dying wave. It spreads upwards to the prefrontal cortex. From there it returns to where it started, creating a resonant circuit of activity. Such is the diffusion of information that occurs when we are aware of it. It reaches the *global work area,* where it becomes available to other regions of the brain.

On the other hand, as Kandel[36] adds, unconscious information processing occurs simultaneously in many different areas, but that information is not sent to other regions. As we read these words, for example, we are aware of our environment: the ambient sound, temperature, humidity, light levels, and so on. This sensory information about the environment is processed unconsciously in the brain, but because it is not widely disseminated, we are not aware of it while we are reading. Experiments have shown that information can enter the brain *without conscious perception*. However, this information *can affect behaviour*. This is because unconscious *cerebration* is not limited to sensory information. While the simple recognition of a word occurs unconsciously, its meaning is accessed at much higher levels of brain processing without our

[35] Kandel, E R, *La nueva biología de la mente. What brain disorders tell us about ourselves*, Paidós. Barcelona, 2022, p. 239.

[36] Op Cit, pp. 241-242.

being aware of it at all. It seems that the way to affect behaviour would be through the *adaptive unconscious*, an idea introduced by the cognitive psychologist Timothy Wilson. The biological function of the adaptive unconscious in decision-making was discovered by Libet, to whom we have referred above and to which we cannot devote more attention in this brief consideration.

During the course of human evolution, as Bargh[37] notes, our basic psychological and behavioural systems were originally unconscious and existed before the late emergence of language and the conscious and intentional use of those systems. The fundamental instinct for physical safety is a powerful legacy of our evolutionary past and exerts a constant influence, responding to modern life often in surprising ways, such as influencing political voting. The right region of the amygdala – the neural headquarters of fear – is larger in people who identify as politically conservative. In lab tasks that involved taking risks, this fear centre of the brain is activated much more in those who declare themselves Republicans than in those who declare themselves as Democrats. So there is a connection between the strength of unconscious motivation for physical safety and a person's political attitudes. And research has shown that progressives can be made more conservative by threatening them and provoking fear.

It is clear that these fears can be first identified today on an individual scale just by analysing in detail the surfing data of millions of citizens and processing them to identify their psychological tendencies and then manipulating them in both subliminal and direct ways, as we shall see below.

The ease with which something comes to mind is called the "*availability heuristic."* The frequency with which something like an image or set of images is used can be accurately induced through various information vectors. The availability heuristic was discovered by Daniel Khaneman and Amos Tversky, as argued by Bargh[38]. These frequency judgements matter in our daily lives because we make decisions based on how often various things happen or are likely to happen.

Behaviour is an unconscious and involuntary effect of the emotional state or states that are processed in the anterior insular cortex or insula as Kandel (248: 2022)[39] points out, which is a small island located between the parietal and temporal lobes. The insula reflects feelings; it is the awareness of physiological reactions to emotional stimuli. The insula not only evalu-

[37] Bargh, John, *Unaware. The power of the unconscious to discover why we do what we do*, Penguin, Barcelona, 2023, p. 52.

[38] Op Cit, p.163.

[39] Op Cit, *The New Biology of the Mind. What brain disorders tell us about ourselves*, p. 344.

ates and integrates the emotional or motivational significance of these stimuli but also coordinates external sensory information and internal motivational states. This awareness of physiological states is a measure of self-awareness. There is evidence, as Bargh[40] points out, that compulsive shoppers are often depressed and that shopping helps them feel happier (or at least not as sad). That sadness underlies a large proportion of compulsive shopping, as proven by the fact that antidepressants are effective in reducing compulsive shopping behaviour. Not to mention that sadness also predisposes people to pay more for the same products.

Today it is possible to know the emotional state of people, for example, with precision neuro-technological *headsets* such as *Kernel* units[41] that read mental states, being this information very valuable for commercial organisations. Well, these would be areas where the AI we are considering should not have access if the data parameterisation is outside the sphere of medical data, which is the case with devices such as NeoRythm or Muse 2 among many other devices that are not regulated by strict medical data protection regulations and in which neurobiological data of users are made available and processed by multinationals outside the EU and under the regulations of *devices for recreational use*, which represents a risk of the utmost importance for the mental privacy of citizens and which is included for the first time in the Leon Declaration on European neurotechnology.[42]

But there are many other forms of unconscious manipulation, and it is important in this regard to note Zajonc's research on the *mere exposure* effect, which was very relevant for several reasons (Bargh)[43]. Firstly, he showed how we can develop tastes and preferences unconsciously, without intending to, just according to the frequency of an experience. Zajonc argued that we often show immediate affective reactions to stimuli such as paintings, sunsets, food, or other people without thinking carefully about it first, what Russell Facio would later call "*automatic attitudes,"* later identified as the paradigm of *affective priming*. A later study by Chris Fritch and colleagues at University College of London concluded that our brains store our current behavioural intentions in areas of the prefrontal and premotor cortex, but the areas used to guide that behaviour are located in a different anatomical area of the brain: the

[40] Op Cit, *Without realising it. The power of the unconscious to discover why we do what we do*, pp. 156-159.

[41] https://www.kernel.com/

[42] Leon Declaration on European Neurotechnology: A people-centred and human rights-based approach. Informal Meeting of Telecommunications Ministers, Leon 23-24 October 2023, p. 2.

[43] Op Cit, *Without realising it. The power of the unconscious to discover why we do what we do*, p. 179.

parietal cortex. This discovery helps explain how *priming* and other unconscious influences can affect our behaviour. *Priming* and external influences on our behaviour can activate guided behaviour – motivated thinking – in one part of the brain *independently of the intention* to perform that behaviour, which is located in a very different part of the brain. It seems that William James was right when, in his famous chapter on "The Will," he argued that our behaviour actually arises from unconscious and unintentional sources, including behaviours appropriate to and suggested by what we are seeing and experiencing at any given moment in our world. Our conscious acts of will, James said, are acts of control over those unconscious impulses, allowing some to manifest and others not.

The human mind is a kind of mirror: it generates potential behaviours that reflect the situations and circumstances of the environment in which we find ourselves: a glass of water says "drink me," a flower says "water me," a bed says "lie down," and museums say "admire me."We are all programmed in this way, to react to external stimuli. Without us realising it, what we see is what we do. This is very important to substantially alter the behaviour of a person or a group of people in a significant way by making decisions that they would not have made *without manipulation*. When behavioural modifications are introduced unconsciously through *priming* and through different reinforcing means to try to modify social ideas, it is possible to combine them with the behavioural guidance patterns of Overton's window on a social scale[44] . The human brain is very sensitive to unconscious manipulation for reasons that, as we can see, are purely neurobiological. To conclude this section, let us recall the neuroscientific research on the brain's motivational circuits carried out by Mathias Pessiglione and Chris Fritch, Bargh,[45] who have confirmed

[44] The Overton window is a model for understanding how ideas in society change or are intentionally changed by *power lobbies* over time and influence policy. The central concept is that politicians are limited in the policy ideas they can support; they generally only pursue policies that are widely accepted throughout society as legitimate policy options. These policies fall within the Overton Window. Other policy ideas exist, but politicians risk losing popular support if they advocate these ideas. These policies are outside the Overton Window. But the Overton Window can change and expand, either increasing or decreasing the number of ideas that politicians can support without unduly risking their electoral support. Occasionally, politicians can influence the Overton window by boldly endorsing policies beyond its boundaries, but this is a rare occurrence. More often, the window moves because of a much more complex and dynamic phenomenon, which is not easy to control from above: the evolution of social values and norms, although it will be the large power groups that support, for example, by *primacy*, the new ideas that they want to support for multiple purposes. More details can be found at: https://www.mackinac.org/overtonwindow (viewed 10 October 2023).

[45] Op Cit, *Without realising it. The power of the unconscious to discover why we do what we do*, p. 289.

that the perception of a reward activates the brain's reward centres whether or not the person consciously perceives the external reward. Participants performed better on the task at hand when the subliminal image of a pound coin (the reward for doing the activity well) appeared before the task, as opposed to when the image of a penny appeared before the task. Furthermore, the brain's reward centre in the hindbrain was more active in the pound condition than in the penny condition. As Dehaene concludes[46], our brain has a set of *intelligent unconscious devices* that constantly monitor the world around us and assign values to it that guide our attention and shape our thinking. Thanks to these subliminal labels, the amorphous stimuli that bombard us become a landscape of opportunities carefully ordered according to relevance to our current goals. Below our level of awareness, our unconscious brain assesses latent opportunities at all times, attesting to the fact that our attention operates largely subliminally.

IV. An Artificial Intelligence that processes everything

Let us assume with Metzinger[47] that the neural correlate of the conscious experience that accompanies *deliberate lying or any other kind of unconscious thought* could be identified (in fact, early candidates are already under discussion). From there we could build efficient high-tech detectors that no longer rely on superficial psychological effects, such as capillary electrical conductivity and changes in peripheral blood flow. This could be an extremely useful tool in the fight against crime and terrorism while fundamentally changing our social world. Something that had hitherto been the paradigm of privacy – the contents of the mind – would suddenly become a public matter. The simplest forms of political resistance, such as confusing the authorities during interrogations, will disappear. On the other hand, society will benefit from increased transparency in many ways. Innocent prisoners could be spared their sentences. Imagine that during the presidential campaign debates, a red light would go on in front of one of the candidates every time the neural correlate of lying was activated in his or her brain. But virtually infallible lie detection would do more than that, it would change our self-models. If we, as citizens, knew that in principle secrets no longer exist, that we can no longer withhold information from the state, one of the pillars of everyday life (at least in the

[46] Dehaene, Stanislas, *La conciencia en el cerebro. Descifrando el enigma de cómo el cerebro elabora nuestros pensamientos*, Siglo XXI editores, Argentina, 2015, pp. 106-107.

[47] Op Cit, *The tunnel of the self. Science of the mind and myth of the subject*, p. 314.

West), the enjoyment of intellectual autonomy and freedom of thought, we add, would disappear. The mere perception of the existence of such neuroforensic technologies would be enough to bring about change. Would we want to live in such a society?

We say freedom of thought because this freedom supposes an infinity of mental options that, although they never materialise, are part of our mental rehearsals or pre-models[48] of behavioural action in an infinity of contexts, but not being materialised does not mean that they have not been previously thought and discarded, but what would happen if these thoughts – not concluded – could be recorded and known by mental recording technologies?

It is feasible to think that a system of transmission of all this enormous amount of information that needs to be monitored both in its input and output to the brain could take advantage of technologies already in use but adapt them to the needs of neurotechnologies, Thus, it would not be out of the question that in a few decades our mobile phones could be implanted in our brains, and with high-bandwidth technologies such as 6G, we would be permanently connected, and neuroprostheses would become independent of equipment such as laptops or medical centres.Oligopolistic macro corporations such as Google, Facebook, Meta, Microsoft, Amazon would filter all our highly sensitive information by means of AI.

Options that were impossible and unthinkable decades ago are now beginning to be seen as possible, opening up many essential questions. What happens if Alzheimer's patients are implanted with *different memories* from those they originally developed in their lives when their brains processed the information that shaped their *pre-illness* identity? They would be the same person, but their identity would have changed. Ensuring the integrity of their memory, their memories, and their mental privacy are major challenges. *Neuro-rights* are intended to advance a preventive legal response to a vast array of ethical dilemmas that we will face sooner rather than later, the prevention of invasive techniques that can already be seen today as contradictory to "classical" human rights in many dimensions and that will pose spectacular challenges in the evolution of our own species from the moment we know how to modify the content of our minds – the neurobiological correlates – on the basis of an increasingly detailed knowledge of

[48] People fantasise and project in their imagination a multitude of neuro-virtual realities that they will never realise but which, even if they are not realised, remain because they are processed in the memory: suicidal ideations, morally or socially reprehensible thoughts of a sexual nature with other people if they are externalised or even thoughts of a criminal nature that could be accessed in the future and be evaluated or judged also by those internal thoughts.

the neurobiological architecture of a central organ in our lives, such as the brain, is approaching.

Neurotechnologies, as pointed out by Müller and Rotter[49] , open up a new world that must be shaped with new ideas and new tools; they represent an opportunity and a challenge that must be faced without fear, but under two essential principles that must operate simultaneously: *the precautionary principle*[50] *and the responsibility principle.* The precautionary principle only applies to the idea of a possible risk, even if there isn't enough or clear scientific evidence to fully show, measure, or figure out its effects. Neurotechnology is developing very rapidly, possibly exponentially. But humans, in our process of adapting to events and challenges, are instead *linear*. When *linear* human beings are confronted with *exponential* change, we cannot adapt to that change easily and this is what we call a *paradigm shift* full of derivatives in a multitude of dimensions and areas of life: social, economic, political, legal, emotional, etc.

The principle of responsibility is clear: we must be responsible for any damage that may be caused by neurotechnological applications that are marketed either in the medical or recreational fields, which *must be brought back under* medical evaluation regulations if we want maximum control over their applications; otherwise recreational applications may *lead* to *the* loss of extraordinarily sensitive personal information and data, which we consider unacceptable, at least with the technical model of data protection currently available to us, which we understand is not ideal for the citizen but is ideal for the organisations and multinational companies that use them, especially when such aggregate data can be delocalised in a global world and processed in worldwide locations that have little or no respect for human rights. The company 23andMe was hacked on 9 October 2023, and millions of pieces of data were stolen from its database. 23andMe is a genetic company in charge of analysing DNA samples from millions of patients. This company is dedicated precisely to receiving saliva samples from its clients to carry out genotyping in order to determine which genes are being expressed and which

[49] Müller, O and S Rotter, Neurotechnology: Current Developments and Ethical Issues, *Frontiers in Systems Neuroscience*, December 2017, Vol 11, article 93, pp. 1-4.

[50] Although the TFEU only explicitly mentions the precautionary principle in the environmental field, Art. 191, its scope of application is much broader. This principle covers specific cases where scientific data are insufficient, inconclusive, or uncertain, but where a preliminary objective scientific assessment leads to the suspicion that there are reasonable grounds to fear that potentially dangerous effects on the environment and human, animal or plant health could be incompatible with the high level of protection chosen. "On the use of the precautionary principle", Communication from the Commission, COM (2000) 1 final, Brussels, 2, 2, 2000.

are silenced. Depending on the genes that are expressed in the subject, it can be determined that he or she may be predisposed to different diseases. The genetic data are for sale on the Dark Web at a price ranging from 1 to 10 dollars. Think what it would be like to have access to or steal a person's life experience as encoded by a laboratory and extracted by a BCI. It is clear that data and security policies would not be able to compensate a person for such a loss if these life experiences of all kinds, personal, ideological, and sexual, including images, were subsequently made public. *Mental privacy* requires new *technical* developments, as the current ones are extremely inefficient. Principles 6, 7, 8, and 10 of the global AI Code of Conduct of 30 October 2023, known as the Hiroshima AI process, approved by the G7, are specifically in this sense, although it suffers from the fact that it is a voluntary international code of conduct for companies. We believe in and have long shared the positive idea put forward by Acemoglu and others[51] regarding the creation of a *data market* in which every citizen would have access to the information that each company has about them and would receive a share of the income generated, i.e., a share of the revenues generated, [52] of personal data, since if the fuel of BigData and AI is data, whoever controls the data will, to a large extent, control AI, for example, by forcing the *exclusion of* data owned by someone through criminal law if their property rights are violated.

The use of neurotechnologies directly addresses the problems we have considered and many others of a similar nature in what Sadin[53] calls the *anthropological condition* that increasingly intertwines human and artificial organisms with fundamental ethical problems that the law cannot ignore, as Lenca and Andorno[54] point out from the perspective of such neuro-rights. Here we can only give a brief outline of how the first attempts at regulation are developing. Neuro-rights would form part of the fourth generation rights and have to do with legal goods affected by AI, genetics, and bioengineering, as well as the whole set of neurosciences that affect the *free development of personality* (Art. 10.1 Spanish Constitution); *physical and moral integrity (*Art. 15 Spanish Constitution); *the right not to be forced to declare ideology, religion, or beliefs* (Art. 16.2

[51] D. Acemoglu, *The impact of Artificial Intelligence will be a mixture of the printing press, the steam engine and the atomic bomb*, at: https://shapingwork.mit.edu/news/daron-acemoglu-the-impact-of-artificial-intelligence-will-be-a-mix-of-the-printing-press-the-steam-engine-and-the-atomic-bomb/ (Visualised, October 2023).

[52] https://www.elnotario.es/opinion/opinion/743-patrimonializar-los-datos-de-caracter-personal-argumentos-para-un-debate-0-022018592825176746

[53] Sadin, E, *La humanidad aumentada*, Caja Negra, Buenos Aires, 2017, p. 152.

[54] Lenca M and R Andorno, Towards new human rights in the age of neuroscience and neurotechnology, *Life Sciences, Society and Policy*, 2017, pp. 5-13.

Spanish Constitution); or *personal privacy* (Art. 18.1 Spanish Constitution) in the case of our constitutional legal system.

V. Addictive technologies. AI action on groups, minors, young people, and other groups.

The fact that China[55] has developed legislation to limit the use of games by children to a maximum of *three hours per week* also responds to the public authorities' awareness that these tools generate very severe *subliminal addictive problems* in the development of minors with serious pathologies for their normal emotional and intellectual development. In fact, according to the WHO, this addiction has been included in the recent international classification of mental illnesses, CIE-11 and, specifically, it is listed as 6C51.0 Gaming Disorder[56] .

As Echeburúa points out[57] , connecting to the Internet as soon as they wake up, when they get home, or just before going to bed and thereby reducing the time devoted to daily tasks (eating, sleeping, studying, or chatting with the family) are some of the usual behaviours of those addicted to social networks or, where appropriate, to new technologies or video games. In other words, more than the number of hours, the determining factor in addiction *is the degree of negative interference* that this behaviour exerts on the daily life of

[55] China bans minors from spending more than three hours a week playing online games. The announcement comes amid growing concern among authorities about addiction to the activity, which they have called "spiritual opium" https://elpais.com/tecnologia/2021-08-30/china-limita-a-tres-horas-semanales-la-practica-de-juegos-online-por-parte-de-los-menores.html (viewed 23 Sep 2023).

[56] *Description*: Video game use disorder is characterised by a pattern of persistent or recurrent gaming behaviour ("digital gaming" or "video gaming"), which may be online (i.e., internet) or offline, and is manifested by: 1. impaired control over gaming (e.g., onset, frequency, intensity, duration, termination, context); 2. increased priority given to gaming to the degree that it takes precedence over other interests and activities of daily living; and 3. continued or increased gaming despite it having negative consequences. The pattern of gambling behaviour may be continuous or episodic and recurrent. The pattern of gambling behaviour results in marked distress or significant impairment in personal, family, social, educational, occupational, or other important areas of functioning. The gambling behaviour and other features are usually evident for a period of at least 12 months for a diagnosis to be assigned, although the required duration may be shortened if all diagnostic requirements are met and symptoms are severe. https://icd.who.int/browse11/l-m/es#/ http%3a%2f%2fid.who.int%2ficd%2fentity%2f1448597234 (viewed 09/2023).

[57] Echeburúa, E, Addicted to new technologies, *Investigación y Ciencia, (Mente & Cerebro)*, May-August, 2013, pp. 36-37.

the person affected. Thus, the smartphone – probably already smarter than the average user – creates dependency in younger individuals, who consider the device indispensable for life and do not know when to do without it (Haidt[58]). Their attention to the messages they receive is constant, so that they often neglect other important activities, including *face-to-face communication*, to reply to virtual contacts or what has come to be called *asynchronous communication* because of the *stress* caused to young people by the fear of real-time communications[59]. The consequences of smartphone abuse also involve a variety of negative effects: there is an attentional focus on the device and its applications, physical activity is reduced, and one is not able to diversify one's time and take an interest in other activities or topics. The subject shows anxiety about social networks and there is a flow of *transreality* reminiscent of the addictive experience of drugs. A snowball effect is created, as problems spread to all personal areas (health, family, school, and social relationships). In short, *dependence* and the subordination of lifestyle to the maintenance of the habit form the core of addiction. Thus, dependence on social networks is not so much characterised by the type of behaviour involved but by the form of relationship that the subject establishes with it. All this can lead to a kind of *relational illiteracy* and facilitate the construction of *fictitious, deficient, and defective social relationships.*

The indiscriminate and uncontrolled use of networks by minors – as in the case of the game already considered – can generate serious mental health problems in young people. As recently revealed by *The Wall Street Journal,*[60] Facebook, the company that owns Instagram, is clearly aware that Instagram is a *toxic* application for teenagers. The use of the application by millions of young people around the world generates a major mental health problem, as a large part of them, but particularly those under 22 years of age, are addicted to it, which the company minimises to the public. Young women who have grown up on this social network are especially vulnerable because their friends who use it at that age have manipulated them into emotional dependence. The newspaper cited internal Facebook studies over the past

[58] Haidt, J, *La generación ansiosa. Por qué las redes sociales están causando una epidemia de enfermedades mentales entre nuestros jóvenes*, Deusto, Barcelona, 2024.

[59] Don't call; send an audio: millennials no longer talk on the phone out of stress. In the era of post-text and frenetic information consumption, asynchronous communication (i.e., fragmented conversation) is gaining ground. Calling is perceived as almost invasive. https://elpais.com/ideas/2021-11-21/no-llames-manda-un-audio-los-mileniales-ya-no-hablan-por-telefono-por-estres.html (viewed on 22-9-2023).

[60] https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739

three years that examined how Instagram affects its young user base. An internal Facebook presentation noted that among teens who reported suicidal thoughts, 13% of users were British and 6% of US users traced the issue of suicide to Instagram.

TikTok gained 682 million new users last year, each of whom spends an average of 50 minutes a day on the Koetsier app[61] but what makes it so addictive? It simply puts the user *in a pleasurable state of mind,* generating dopamine, by visualising and watching these images and getting carried away (Haynes[62]). It is almost hypnotic; you will keep looking and looking. When you scroll from one image to another, sometimes you see a picture or something that is striking, attractive, and catches your attention. At that moment, you get that little dopamine surge in the brain in the pleasure centre of the brain the *nucleus accumbens.* So you want to keep scrolling, browsing, and viewing images through these endless networks of images. You keep scrolling because sometimes you visualise something that is pleasurable but *sometimes not.* And that *differentiation* is the key, much like a slot machine in a casino, as James and colleagues point out.[63] Platforms like TikTok, Instagram, Snapchat, and Facebook have adopted the same psychological principles that have made gambling addictive. In the federal lawsuit – Case *4:23-cv-05448*[64]– filed in the Northern District of California by 33 General Attorneys, the States allege that Meta products have harmed minors and contributed *to a serious mental health crisis in the United States* based on the arguments summarily considered here. The multi-state federal lawsuit on Tuesday, 24 October 2023, involves California, Colorado, Connecticut, Delaware, Georgia, Hawaii, Idaho, Illinois, Indiana, Kansas, Kentucky, Louisiana, Maine, Maryland, Michigan, Minnesota, Missouri, Nebraska, New Jersey, New York, North Carolina, North Dakota, Ohio, Oregon, Pennsylvania, Rhode Island, South Carolina, South Dakota, Virginia, Washington, West Virginia, and Wisconsin. It should be noted that in Meta's internal studies, the multinational was aware that its social networks had and still have an addictive capacity on users, as Horwitz argues[65] .

[61] https://www.forbes.com/sites/johnkoetsier/2020/01/18/digital-crack-cocaine-the-science-behind-tiktoks-success/?sh=40498c0678be (viewed on 10 September 2023).

[62] T Haynes, *Dopamine, Smartphone & You: A Battle for Your Time*, Harv. Univ. SITN Blog, May 1, 2018, https://archive.ph/9MMhY

[63] J RJE, O'Malley C and Tunney RJ, Why are Some Games More Addictive than Others: The Effects of Timing and Payoff on Perseverance in a Slot Machine Game. Front. Psychol. 7:46. (2016) doi: 10.3389/fpsyg.2016.00046.

[64] https://ag.ny.gov/sites/default/files/court-filings/meta-multistate-complaint.pdf

[65] Horwitz J, *Broken code. Manipulación política, fake news, desinformación y salud pública*, Ariel, Barcelona, 2024.

In psychological terms, it is called "*random reinforcement*" and means that sometimes you win and sometimes you lose. Studies on *random reinforcement* come from research with rats in feeding and reward processes. And that is how these platforms are – consciously designed; they are exactly like a gambling machine. We are aware that gambling addiction exists, and as an extreme pathology, pathological gambling. But we don't often consider how our "*smartphones*" and these widespread apps have these same addictive qualities built into what they offer us *by design*, i.e., clearly addictive technologies are consciously offered to users. That is the reason behind the demand.

The discovery of this technique by Morgan[66] came from a set of experiments in which the response of rats to conditioning was studied. Through specific conditioning, rats were given a reward each time they performed a simple task. Rewarding is a good way to encourage someone to do something. But the experimenters realised that there was *something better* than "*positive reinforcement.".* It was "*random reinforcement.".* Under random reinforcement, when the rat performed the task*, sometimes it got a reward and sometimes it did not.* Although it may seem counterintuitive that rewarding someone always or every time will achieve a goal that you want the subject to perform, the truth is that it is much more effective *to reward only sometimes.* In the case of human beings, it is emotions that are the raw material with which these electronic conditioning systems play. The emotion is what is really relevant and it is the "*uncertainty"* that triggers the exciting and addictive feeling that achieving the goal triggers the *release of* Eimeren *dopamine.*[67] If our football or basketball team wins every game, we soon stop caring. The person who is always on top of us, attending to us and giving us all the attention all the time, starts to annoy us. In fact, we are always attracted to that which seems most difficult to achieve. If there is always gratification, we soon reach a *level of saturation* and become conditioned in such a way that the reward no longer produces the same level of well-being and pleasure as the first time. We value things according to what it costs us to get them, so we value what we have or can get "*sometimes, when we win*" more than what we always have.

We may think we are much more sophisticated than rats, but for this particular aspect of human behaviour, this is not the case; we operate like them and many other animals. The *excitement* of live matches, the fact that we are

[66] Morgan, M. J., Effects of random reinforcement sequences, *Journal of the Experimental analysis of behaviour,* No. 2, (September), 22, 1974.

[67] Eimeren, TV, Temporarily giving up technological devices, so-called "dopamine fasting", can prevent addiction to these devices, in *Mind & Brain*, November-December, no. 111, 2021, pp. 66-67.

passionate about games of all kinds of gambling or chance, looking for prizes that we may *or may not get*, the fact that we have all wanted what we cannot achieve and do not value what we always have or have already achieved. Each and every one of these very human facts is the empirical demonstration of how *random reinforcement* is the most addictive and it is precisely what these platforms offer people; naturally, the impact on youth is quantitatively and qualitatively much more addictive than on adults because their brains are not yet developed. But among the groups that can be subject to manipulation, there are also groups such as adults affected by cyber gambling, groups that must be protected.

VI. Conclusions

Kant pointed out that *an individual's intelligence is measured by the amount of uncertainty he or she is able to withstand.* Almost a hundred years later, F. Scott Fitzgerald coined one of the most famous definitions of intelligence in history: *the test of a first-rate intellect is its ability to handle two conflicting ideas at the same time and still maintain the ability to function.* In VUCA environments (volatility, uncertainty, complexity, and ambiguity), such as AI specifically, there are truly useful advantages for the whole of society and very important risks that oscillate at the same time on a horizon of indeterminacy that consolidates step by step depending on the social decisions we make about them. The regulation of AI in the European Union, the Artificial Intelligence Act,is a happy temporary crystallisation of a regulation that favours the social advantages of AI and minimises its risks; in this sense, there are reasons to consider it a good *initial* regulation, particularly in the prevention of unacceptable risks of mental manipulation, which are the ones we have dealt with in this chapter, without prejudice to the fact that we must remain attentive to the development of the regulation as it fits in with the technological reality that is often resistant to being regulated, but that is the vocation of law and perhaps its only truly consistent merit.

THE REMAINING ARTIFICIAL INTELLIGENCE SYSTEMS PROHIBITED OR UNACCEPTABLE IN THE ARTIFICIAL INTELLIGENCE ACT

Pere Simón Castellano[1]

Senior Lecturer in Constitutional Law at the International University of La Rioja – UNIR

I. Introduction

Article 5 of the AIA expressly prohibits the placing on the market, putting into service or use of certain AI systems and certain uses and "practices". Indeed, Chapter II of the AIA is precisely entitled "Prohibited AI practices". Much of the scope of the prohibition of certain AI practices has been the subject of study in the chapters that have preceded *ut supra* to the one that the reader now has the pleasure of writing.

Other studies in this work on prohibited AI focus on recognition through biometrics; AI systems that use subliminal techniques that transcend a person's consciousness or deliberately manipulative or deceptive techniques with the purpose or effect of substantially altering the behaviour of a person or a group of persons, appreciably impairing their ability to make an informed decision and causing a person to make a decision that they would not otherwise have made, in a way that causes, or is likely to cause, significant harm to that person, another person or a group of persons. The prohibition of systems which exploit vulnerabilities of a person or a specific group of persons resulting from age or disability, or from a specific social or economic situation, with the purpose or effect of materially altering the behaviour of that person or a person belonging to that group in a way that causes, or is reasonably likely to cause, significant harm to that person or another person has also been considered; in AI systems for the purpose of assessing or ranking individuals or groups of individuals over a given period of time on the basis of their social behaviour or known, inferred or predicted personal or personality characteristics, such that the resulting citizen score results in detrimental or unfavourable treatment. It should also be remembered that risk assessments of natural persons and the commission of criminal offences are analysed in the study by Fernan-

[1] This work is carried out in the framework of the Project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by MCIN/AEI/10.13039/501100011033/ ERDF, EU.

do Miró Llinares and Mario Santisteban Galarza, in the thematic block on high-risk AI systems.

However, the AIA foresees other prohibited practices or unacceptable AI that are the subject of this study: prohibition of AI based on facial recognition with images extracted from the Internet or television; emotion recognition AI systems; biometric categorisation systems that individually classify individuals on the basis of their biometric data.

The issue is important because in practice and at the international level, AI-based systems have already started to be implemented, which has generated controversy and sometimes even rejection. Beyond the traditional use to identify individuals, advanced biometric systems can now recognise emotions, classify people, detect behaviour and thoughts, and even assess personalities, including so-called intelligent polygraphs. The implementation of these systems in border control is not a recent practice; they were introduced in Arizona more than a decade ago. More recently, in the United States, the Automated Virtual Agent for Truth Assessment in Real Time (AVATAR) has been used to analyse both verbal and non-verbal behaviour of travellers, and has also been tested at Bucharest airport, among others. The European Commission funded the *Intelligent Portable Control System* (iBorderCtrl) project[2], which uses tools for deception detection and risk-based assessments, which has provoked a remarkable reaction from civil society, including a European citizens' initiative and the *reclaimyourface.eu* campaign.

Another example can be found in Brazil. On 7 May 2021 the São Paulo Court of Justice banned the São Paulo Metro concessionaire from using the "Digital Interactive Door System" (DID) with facial recognition, the system inferred emotions, gender and age of people to personalise advertising[3].

Large companies and platforms also have biometric and facial recognition systems not only to identify people but also to detect emotions, moods, etc. In June 2022, Microsoft announced that it was withdrawing its *Azure Face* systems[4], having previously stopped selling this type of technology to the US police. Meta-Facebook has had emotion recognition patents since 2017[5]. In November 2021 it phased out the controversial use of facial recognition[6].

[2] On border use, see Sánchez Monedero, J. and Dencik, L. "The Politics of Deceptive Borders: "Biomarkers of Deceit" and the Case of iBorderCtrl"". *Information, Communication & Society*, vol. 1, 2020. https://doi.org/10.1080/136911

[3] https://www.accessnow.org/sao-paulo-court-bans-facial-recognition-cameras-in-metro/

[4] https://azure.microsoft.com/es-es/products/cognitive-services/face/

[5] https://www.cbinsights.com/research/facebook-emotion-patents-analysis/

[6] https://about.fb.com/news/2021/11/update-on-use-of-face-recognition/

However, the reader should note that the AIA contains specific rules for the protection of individuals in relation to the processing of personal data that restrict the use of AI systems for remote biometric identification for law enforcement purposes, the use of AI systems for carrying out risk assessments of natural persons for law enforcement purposes and the use of AI systems for biometric categorisation for law enforcement purposes. This is why the AIA, as far as these specific rules are concerned, finds its basis and rationale in Article 16 TFEU. In the light of these specific rules and the recourse to Article 16 TFEU, we have to take into account in particular the opinion of the European Data Protection Committee in this respect.

II. Definitions and terminology: biometric categorisation technologies, emotions, biometric data and facial recognition

Automatic or automated facial recognition has to be included in a whole range of "biometric techniques" ("biometric identification", "biometric categorisation", "behaviour detection", "emotion recognition", "biometric data" processing, "biometric profiling", etc.).

Biometric identification systems have been defined as automated processes used to recognise an individual by measuring, storing and comparing biometric data relating to physical, physiological or behavioural characteristics[7]. These biometric characteristics are universal -all humans have them-, singular and unique, and invariant throughout life.

Biometric data processing systems are based on collecting and processing personal data relating to the physical, physiological or behavioural characteristics of natural persons, including, as has recently become apparent, their neural characteristics, by means of devices or sensors, creating biometric templates (also called signatures or patterns) that enable the identification, tracking or profiling of such persons (i.e., 'processing', Art. 4.2 of the GDPR). The GDPR defines in Article 4.14 biometric data as "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data", and the definition states that biometric data are all data that allow the identification or authentication of a person.

[7] See the compilation of related concepts available in Gallego Rodríguez, P. "Los registros biométricos y su aplicación al proceso penal desde una perspectiva constitucional", in Calaza López, S. and Llorente Sánchez-Arjona, M. (dirs.), *Inteligencia artificial legal y administración de justicia. Aranzadi*, Cízur Menor, 2022, pp. 211-255, see pp. 234 et seq.

However, biometric data may allow the authentication, identification or categorisation of natural persons and the recognition of emotions of natural persons. We say "may" because, as will be seen, the final version of the AIA does not retain this reference to biometric data enabling the identification or authentication of a person. Biometric data have the same definition in Art. 3.13 of Directive (EU) 2016/680[8], Article 4.14 of the GDPR and, except for the reference to unique identification, also in the AIA, in Art. 3.34. These are 'personal data obtained from specific technical processing, relating to the physical, physiological or behavioural characteristics of a natural person, such as facial images or dactyloscopic data'. In the latest approved version, the reference to "allowing or confirming the unique identification of that person" has been deleted from the original text of Article 3.33 of the AIA. The reason behind the exclusion of this mention in the AIA is not because of redundancy insofar as only personal data are personal data that allow for the identification of a person, and it follows that biometric data are not biometric data that do not allow for the unique identification of a person. The reason is that it is appropriate to exclude systems of mere biometric verification, which includes authentication, whose sole purpose is to confirm that a specific individual is the person they claim to be, as well as the identity of an individual for the exclusive purpose of granting access to a service, unlocking a device, or providing security access to a location. Those systems are not prohibited as long as the risk is, for obvious reasons, withing an acceptable or tolerable threshold.

Recital 15 of the AIA helps to delimit the content by stating that the concept of "biometric identification" referred to in this Regulation should be defined as the "automated recognition of physical, physiological and behavioural human features such as the face, eye movement, body shape, voice, prosody, gait, posture, heart rate, blood pressure, odour, keystrokes characteristics, for the purpose of establishing an individual's identity by comparing biometric data of that individual to stored biometric data of individuals in a reference database, irrespective of whether the individual has given its consent or not. This excludes AI systems intended to be used for biometric verification, which includes authentication, whose sole purpose is to confirm that a specific natural person is the person he or she claims to be and to confirm the

[8] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data. Hereinafter referred to as Directive (EU) 2016/680.

identity of a natural person for the sole purpose of having access to a service, unlocking a device or having security access to premises."

It is also essential to pay attention to the two definitions in Articles 3.35 and 3.36 of the AIA, which state that biometric identification should be considered as "the automated recognition of physical, physiological, behavioural, or psychological human features for the purpose of establishing the identity of a natural person by comparing biometric data of that individual to biometric data of individuals stored in a database" and biometric verification should be understood as "the automated, one-to-one verification, including authentication, of the identity of natural persons by comparing their biometric data to previously provided biometric data".

A biometric data contained in a system is stored in the form of a biometric template or pattern. A biometric template is a way of writing a human biometric characteristic, such as a face or a fingerprint, in a way that is interpretable by a machine efficiently and effectively for a given purpose or purposes. The biometric template is not intended to be interpreted by a person, such as a photograph, but is intended to be processed in an automated process, i.e., to be efficiently and effectively interpretable by a machine. This form of storage would make it possible to single out an individual and execute actions automatically, profile or infer information about a subject such as attitudes or behaviour patterns, etc.

In the case of identification or authentication operations, for a biometric template to be effective, it is necessary that the templates generated from two different individuals are clearly distinguishable. In this case, the template acts as a unique identifier of the person. The fact that the original face cannot be reconstructed from a biometric template, e.g., from facial recognition, is irrelevant, as it is a unique identifier that uniquely singles out the original face, at least in the context of automated processing. Similarly, a name or a face cannot be reconstructed from the ID number alone. Both unique identifiers, biometric template or ID card number, can be associated with additional personal data and attributes in a file. Unlike an ID number, the biometric template is not assigned to a person, but is generated directly from the observation of unique and unalterable physical characteristics of the individual himself, without the need to rely on documents, other devices or third party databases.

We thus speak of identifiers based on physical, physiological or behavioural characteristics. A distinction is made between "strong" identifiers, which are especially used with the first generation of identification technologies (fingerprints, DNA, iris structure, faces, voice) and "weak" identifiers, which are becoming increasingly important (gait patterns, blood vessel patterns, keystrokes, etc.). The new generation of technologies goes beyond the purpose of identification and is referred to as "behavioural biometrics" for profiling, emotion recognition or categorisation of persons. This is why, as opposed to biometric data linked only to identification, the broader and more inclusive concept of "biometric-based data" is proposed[9].

The issue is important because only biometric data for identification are specially protected data under the special regime of Article 9 GDPR or Art. 10 Directive (EU) 2016/680. The European Data Protection Committee remains clear that, if not linked to identification, they are not sensitive data[10]. This fits with the definition included in the AIA which excludes mere biometric verification systems, comprising authentication, the sole purpose of which is to confirm that a specific natural person is the person he or she claims to be, as well as the identity of a natural person for the sole purpose of accessing a service, unlocking a device, or having security access to premises. It is striking that data revealing our emotions, thoughts, and intentions are not included in the special categories of data[11]. It is also appropriate to deal with the concept of 'biometric inferences' concerning the conclusions or results of the permanent or long-term processing of such biometric-based data.

[9] Parliament's amendments of June 2023 incorporate the concept, in Article 3.33a, of 'biometrics-based data', which it defined as personal data resulting from the specific technical processing of physical, physiological or behavioural signals or characteristics of a natural person, such as facial expressions, movements, pulse rate, voice, heart rate or gait, which may or may not allow the unique identification of a natural person. This definition has been subsumed under recital 15 which refers to physical, physiological or behavioural human characteristics, such as face, eye movement, body shape, voice, intonation, gait, posture, heart rate, blood pressure, smell or keystroke characteristics, in order to establish the identity of a person by comparing his biometric data with biometric data of persons stored in a reference database, whether or not the person has given his consent. Excluded from this are systems for mere biometric verification, which includes authentication, the sole purpose of which is to confirm that a specific natural person is the person he or she claims to be, as well as the identity of a natural person for the sole purpose of accessing a service, unlocking a device or gaining security access to premises.

[10] See *Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement, Version 1.0*, 12 May, https://edpb. europa.eu/our-work-tools/documents/public-consultations/2022/ guidelines-052022-use-facial-recognition_en

[11] On the expansion or insufficiency of the category of specially protected data, as is also the case with health data, see Cotino Hueso, L. "El alcance e interactuación del régimen jurídico de los datos personales y big data relacionados con salud y la investigación biomédica", *Revista de derecho y genoma humano: genética, biotecnología y medicina avanzada*, n.º 52, pp. 57-96, n.º 52 enero-junio 2020.

III. Unacceptable practices: purpose and content of the prohibition

The AIA prohibits the introduction on the market, the placing on the market for this specific purpose or the use of AI systems that create or extend facial recognition databases by non-selectively extracting facial images from the internet or closed-circuit television (CCTV); the introduction on the market, the placing on the market for this specific purpose or the use of AI systems to infer the emotions of a natural person in workplaces and educational institutions, except where the AI system is intended to be installed or placed on the market for medical or security purposes; the introduction on the market, putting into service for this specific purpose or the use of biometric categorisation systems that individually classify natural persons on the basis of their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation; this prohibition does not include the tagging or filtering of lawfully acquired biometric data sets, such as images, based on biometric data or the categorisation of biometric data in the field of law enforcement; the use of "real-time" remote biometric identification systems in publicly accessible areas for law enforcement purposes.

1. Easy recognition via scraping or non-selective extraction of facial images from the Internet and closed-circuit television (CCTV)

Recital 43 of the AIA states that "the placing on the market, the putting into service for that specific purpose, or the use of AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage, should be prohibited because that practice adds to the feeling of mass surveillance and can lead to gross violations of fundamental rights, including the right to privacy".

Facial recognition using *scraping* techniques, also known as non-selective extraction of facial images from the internet and CCTV, is an advanced technology that allows identifying or verifying a person's identity by analysing their facial features. This process is carried out by AI algorithms that extract and process facial data from images available online or captured by surveillance cameras.

Facial *scraping* involves the bulk collection of facial images of individuals from a variety of digital sources, such as social networks, public websites, and surveillance cameras. Unlike authorised data collection, *scraping* is not done with the explicit consent of the individuals whose images are collected. AI algorithms analyse these images to create detailed facial profiles, which can then be used for a variety of applications, from security to personalised advertising.

This method of non-selective extraction raises serious ethical and legal concerns, as it invades the privacy of individuals and may lead to misuse of the information collected. The ability to identify individuals without their consent and in unauthorised contexts poses a significant risk to the fundamental rights of individuals and the free development of their personality, as well as to their personal liberty.

The prohibition of this technique in the AIA finds its core justification, beyond privacy, in the need to prevent abuse, manipulation, discrimination or certain uses linked to cybercrime. The unauthorised collection and use of facial data represents, first and foremost, a direct violation of privacy. The AIA seeks to close the regulatory circle and complement with this prohibition a use that is hardly compatible with the GDPR, protecting European citizens from intrusion into their private lives and ensuring that their biometric data are not used without their knowledge and explicit consent. However, as mentioned above, the essential rationale for this prohibition goes beyond privacy and data protection, and is that images and biometric facial recognition data can be used for discriminatory purposes, such as targeted surveillance of certain ethnic groups or discrimination in access to services and opportunities. By prohibiting these practices, the AIA seeks to prevent abuse of the technology and to ensure fair and equal treatment of all individuals. In addition, the massive collection of facial data without control can lead to the creation of databases susceptible to being hacked or exploited. By establishing clear restrictions, the AIA seeks to ensure that AI technologies are developed and used in a safe and reliable manner, strengthening public confidence in these innovations. It should also be noted that the indiscriminate use of facial *scraping* contravenes fundamental ethical principles, such as respect for human dignity and personal autonomy, and in practice would have harmful effects with a potential detrimental effect on the free development of personality.

We consider the prohibition of facial recognition via scraping in AIA to be a crucial measure to protect the rights and freedoms of individuals, prevent abuse of the technology and encourage ethical and safe development of AI innovations. By setting these limits, the EU is at the forefront in regulating emerging technologies, ensuring a balance between technological progress and respect for fundamental rights.

2. The prohibition of the use of Artificial Intelligence systems to infer emotions in the Regulation

The AIA establishes a clear and strict prohibition on the use of AI systems to infer the emotions of individuals in workplaces and educational establishments, except in situations where such systems are used for medical or security purposes. This provision, included in recital 44 of the AIA, underlines the need to protect the fundamental rights and privacy of individuals from practices that could lead to mass surveillance and serious violations of their privacy.

More specifically, the European legislator tells us that "there are serious concerns about the scientific basis of AI systems aiming to identify or infer emotions, particularly as expression of emotions vary considerably across cultures and situations, and even within a single individual. Among the key shortcomings of such systems are the limited reliability, the lack of specificity and the limited generalisability. Therefore, AI systems identifying or inferring emotions or intentions of natural persons on the basis of their biometric data may lead to discriminatory outcomes and can be intrusive to the rights and freedoms of the concerned persons. Considering the imbalance of power in the context of work or education, combined with the intrusive nature of these systems, such systems could lead to detrimental or unfavourable treatment of certain natural persons or whole groups thereof. Therefore, the placing on the market, the putting into service, or the use of AI systems intended to be used to detect the emotional state of individuals in situations related to the workplace and education should be prohibited. That prohibition should not cover AI systems placed on the market strictly for medical or safety reasons, such as systems intended for therapeutical use".

The ban focuses on AI systems that analyse and determine human emotions through various biometric and behavioural signals. These systems are able to interpret facial expressions, voice tones, gestures and other physical and behavioural characteristics to infer emotional states, such as happiness, sadness, stress, and others. The technology that enables this inference is based on the analysis of large volumes of data, often collected without the explicit consent of the individuals concerned.

The areas of application of the ban include workplaces or workstations and educational establishments. The ban in the workplace responds to concerns that employers may use these technologies to continuously monitor and assess the emotional state of employees. Such a practice could lead to an oppressive work environment and discrimination based on perceived emotions, affecting workers' mental health and privacy. As far as educational institutions are concerned, the use of AI to infer emotions poses significant risks to the privacy and well-being of students. Constant emotional monitoring could interfere with the natural development and autonomy of students, as well as create an environment of surveillance that contravenes the principles of free and open education.

This limits the scope of the prohibition considerably, and indeed, in the AIA itself, the legislator encourages their use in other areas, in particular when referring to medical or security purposes. For example, in a medical setting, such systems may be crucial to diagnose and treat mental or emotional health conditions. Similarly, in security situations, the ability to infer emotions can be vital to prevent immediate threats, such as identifying potentially dangerous behaviour in real time. It is not a *numerus clausus* of exceptions, but it is a closed list of prohibitions, which only projects effects on schools and workplaces.

The ban is based on the protection of the fundamental rights of European citizens, in particular the right to privacy and psychological integrity. Inference of emotions can lead to a form of intrusive surveillance that not only invades people's privacy, but can also manipulate their behaviour and decisions. Moreover, the accuracy of these technologies is not guaranteed and can vary significantly, increasing the risk of errors and misinterpretations.

The introduction of AI systems capable of inferring emotions also raises significant ethical concerns. The potential for these technologies to be used to influence human behaviour surreptitiously raises questions about free will and manipulation. In work and educational contexts, where individuals are already in positions of lesser power compared to employers or educational authorities, this technology could exacerbate power inequalities and lead to abuses. The prohibition of these AI systems in the AIA reflects a firm commitment to protecting human dignity and preventing a surveillance environment that could undermine fundamental rights and freedoms.

3. The use of biometric categorisation Artificial Intelligence systems that individually classify natural persons on the basis of their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical convictions, sex life or sexual orientation.

The AIA also prohibits the use of AI systems for biometric categorisation that individually classify natural persons on the basis of their biometric data. More specifically, the AIA prohibits systems that infer or deduce a natural person's race, political opinions, trade union membership, religious or philosophical convictions, sex life or sexual orientation from their biometric data, such as face or fingerprints.

The object or material scope of the indicated prohibition is specified in the use by AI systems of biometric data to perform personal categorisations. Biometric data includes physical, biological and behavioural characteristics that are unique to each individual, such as fingerprints, facial features, the iris of the eye, among others. These systems can analyse this data to try to infer sensitive information about individuals, such as their race, political beliefs, union affiliation, religious beliefs, sex life or sexual orientation. However, such inferences or predictions from AI systems may be inaccurate or predispose users towards biased decision-making.

The prohibition of such systems in the AIA is therefore justified on several key grounds, mainly focusing on the protection of fundamental rights and privacy of individuals. Deduction or inference or prediction based on sensitive information from biometric data can lead to significant violations of fundamental rights. Privacy, equality and non-discrimination are essential pillars of EU law. Using AI to infer such intimate characteristics may result in misuse of information, discriminatory actions or other forms of social injustice. Furthermore, the use of AI systems for biometric categorisation can contribute to a state of massive surveillance, where individuals are constantly monitored and analysed. This not only infringes on privacy, but also creates an environment of distrust and fear, undermining individual freedom and the right to personal autonomy. Moreover, AI systems are not infallible and can make mistakes in the interpretation of biometric data. If the outcome of the prediction is to make a commercial 'proposal' or a proposal that does not result in an action that has an effect on the rights of data subjects, then it would not project negative or pernicious effects. But incorrect inference[12] from sensitive characteristics can lead to wrong decisions and discrimination. In addition, algorithms can perpetuate and amplify existing biases in the data, resulting in unfair treatment of certain individuals or groups. Finally, the classification of individuals on the basis of their biometric characteristics and the inference of personal and sensitive information also raises ethical questions. Human dignity is compromised when individuals are reduced to a set of biometric data and categorised without their consent.

The prohibition provided for in the AIA is also not absolute, as it establishes or provides for notable exceptions. It does not apply to the lawful tagging, filtering or categorisation of biometric datasets acquired under national or EU law. For example, the classification of images on the basis of hair or eye colour may be permissible in the context of law enforcement, where such data are used for legitimate and specific purposes that do not compromise

[12] They may even derive from the design of the technological solution. Note that correlation refers to the correspondence between two or more actions or phenomena; however, correlation does not imply causation. The output of many AI systems, then, demonstrates a correlation, but not necessarily an effect or consequence, per se. And this may be the source of many misunderstandings or problems regarding the results that AI systems produce. See Lehr, D. and Ohm, P. "Playing with the Data: What Legal Scholars Should Learn About Machine Learning", in University of California Davis Law Review, vol. 51, 2017, p. 671.

the privacy and fundamental rights of individuals. The AIA more specifically states that "this prohibition does not cover the tagging or filtering of lawfully acquired biometric datasets, such as images, based on biometric data or the categorisation of biometric data within the scope of law enforcement".

IV. The coexistence of the regulation foreseen in the AI Act with data protection law: an overlapping and compatible regulation?

It has been pointed out by some authors that the essential defining element of biometric -and facial recognition- systems is their purpose of identifying individuals[13], and this has much to do with the definition of biometric data already reproduced in the GDPR. However, this crashes with the AIA insofar as the latter only provides that biometric data may allow the authentication, identification, or categorisation of natural persons and the recognition of the emotions of natural persons. In the latest approved version, the reference to "enabling or confirming the unique identification of that person" has been removed from the original text of Article 3.33 of the AIA. It excludes systems for mere biometric verification, which includes authentication, the sole purpose of which is to confirm that a specific natural person is who they claim to be, as well as the identity of a natural person for the sole purpose of accessing a service, unlocking a device or gaining security access to premises.

A further distinction should be made, where possible, between biometric identification (*one-to-many*) and biometric authentication or verification (*one-to-one*). This has been done by data protection authorities. As defined by the Article 29 Working Party already in 2012 or by the Spanish Data Protection Agency (AEPD[14]), biometric identification is the process of comparing the biometric data of an individual, acquired at the time of identification, with a set of biometric templates stored in a database of generally identified individuals, i.e., a one-to-many matching process. By contrast, 'one-to-one' biometric 'verification' or 'authentication', i.e., the process of confirming that an individual is who they claim to be by comparing the data only with the identity to be checked, has not traditionally been considered as process-

[13] Cotino Hueso, L. "Reconocimiento facial automatizado y sistemas de identificación biétrica bajo la regulación superpuesta de inteligencia artificial y protección de datos", in Balaguer Callejón, F. and Cotino Hueso, L. (coords.), *Derecho público de la inteligencia artificial, Madrid*, Marcial Pons, 2023, pp. 347,402.

[14] Thus, the Article 29 Working Party since Opinion 3/2012 on the evolution of biometric technologies or the AEPD in its various guides.

ing sensitive data under the special regime of Article 9 of the GDPR. This would be the case, for example, of identification with the personal mobile phone through the fingerprint or face registered on it. This being the case, the current wording of the AIA proposal would fit perfectly with the interpretation that has been made of data protection law. It is not sensitive data and the special regime of Article 9 of the GDPR does not apply, since it is a "one-to-one" verification or authentication, which falls within the exclusion foreseen in the AIA when it tells us that AI systems of mere biometric verification are not affected by the prohibition, which also include authentication, the sole purpose of which is to confirm that a specific natural person is the person he or she claims to be, as well as the identity of a natural person for the sole purpose of accessing a service, unlocking a device or gaining security access to a building, site or premises, shall not be covered by the prohibition.

In these one-to-one verifications, and in a similar sense, the AEPD has indicated that "although it also processes personal data, it does not process the information against a previous database that allows or confirms the identification of individuals one by one"[15]. Therefore, the fact that Article 9.1 GDPR includes among the specially protected data "biometric data intended to uniquely identify a natural person" is not incompatible with the exclusion of the prohibition provided for in the AIA. This is so insofar as it has been interpreted as meaning that the use of biometric data only falls under the specially protected data regime for one-to-many identifications and not for one-to-one identifications.

The European Data Protection Committee has made an interpretative change with its 2022 guidelines[16]. The AEPD has externalised this change of criterion in the Guide to Time and Attendance Processing by means of biometric systems of 23 November 2023, in which it mainly indicates that it should be assumed that both for identification and authentication, we are dealing with a high-risk processing that includes special categories of data. This requires

[15] See AEPD sanctioning procedure PS/00120/2021, p. 28.

[16] See also the position of the European Data Protection Committee, which initially seemed to change its criteria, in *Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement*, Version 1.0, 12 May, https://edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-052022-use-facial-recognition_en. After distinguishing authentication-verification from identification, it states that both cases "constitute a processing of personal data, and more specifically a processing of special categories of personal data". See also the work of Santisteban Galarza, M. "Reconocimiento facial y protección de datos: una respuesta provisional a un problema pendiente". *Revista de Derecho de la UNED* (RDUNED), n.º 28, 2022, pp. 499-526. https://doi.org/10.5944/rduned.28.2021.32887

specific regulatory authorisation. This change of criterion is linked to the recognition and processing of biometric data for the purposes of working time control in the field of employment and not in a specific sector such as identification in the public sector and in particular in the performance of justice. However, it does take into account[17] and for a case such as the present one it includes fully applicable statements: "the requirements of necessity, present in all of them, in addition to those of reservation of law in letters c) and d) and also in the case of letter f) the overcoming of the analysis of prevalence between the legitimate interests of the controller and the interests or fundamental rights and freedoms of the data subject that require the protection of personal data, in particular when the data subject is a child, will have to be met". And that "prior to any decision to implement a time and attendance system using biometric systems, risk management (Art. 24.1 GDPR) and appropriate technical and organisational measures should be implemented by design and by default (Art. 25 GDPR) in order to ensure and be able to demonstrate that the processing is compliant with the GDPR. In particular, in case of high risk, a Data Protection Impact Assessment (DPIA) that includes and also passes the triple test of appropriateness, necessity and strict proportionality set out in art. 35.7.b and also provided for by the doctrine of the Constitutional Court must be passed".

It also concludes that "in the event that the biometric system is implemented with Artificial Intelligence techniques, the prohibitions, limitations and requirements established in the regulations on Artificial Intelligence must be taken into account in order to include them in a processing operation", providing the "minimum default measures", which include "informing the data subjects about the biometric processing; implementing in the biometric system the possibility of revoking the identity link between the biometric template and the natural person; implementing technical means to ensure that the templates cannot be used for any other purpose; using encryption to protect the confidentiality, availability and integrity of the biometric template; using specific data formats or technologies that make interconnection of bio-

[17] pp. 19-20. A limitation on the use of consent is expressed in the same Guidelines in relation to the use of consent in the framework of public authorities:

16. Recital 43 clearly indicates that public authorities are not likely to be able to rely on consent to carry out data processing because when the controller is a public authority, there is always a clear imbalance of power in the relationship between the controller and the data subject. It is also clear in most cases that the data subject will not have realistic alternatives to accept the processing (the conditions of processing) of that controller. The ECDC considers that there are other legal bases that are, in principle, more suitable for data processing by public authorities.

metric databases and unverified data disclosure impossible; deleting biometric data when they are not linked to the purpose for which they were processed; implementing data protection by design; carrying out a Data Protection Impact Assessment prior to the start of processing”.

High-risk Artificial Intelligence systems: delimitation and analysis of certain areas

SCOPE AND DELIMITATION OF HIGH-RISK SYSTEMS IN THE ARTIFICIAL INTELLIGENCE ACT

Lorenzo Cotino Hueso

Professor of Constitutional Law at the University of Valencia. Valgrai

I. High-Risk System Status is Essential for the Regulation

The AIA applies to AI systems,[1] but the fact is that most of the regulation and the imposition of obligations conceived revolves around the AI system being considered as high risk (hereinafter HRS). The HRS status has become the key element of the AIA, mentioned up to 470 times throughout the text and is part of the very "scope of application" of Articles 2, items (2) and (12).

The definition of "high risk" already generated very divergent positions during the process prior to the AIA proposal.[2] The notions that were consid-

[1] cotino@uv.es. OdiseIA. This study is the result of research from the following projects: MICINN Project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by MCIN/AEI/10.13039/501100011033/; "The regulation of digital transformation ..." Generalitat Valenciana "Algorithmic law" (Prometeo/2021/009, 2021-24); "Algorithmic Decisions and the Law: Opening the Black Box" (TED2021-131472A-I00) and "Digital transition of public administrations and Artificial Intelligence" (TED2021-132191B-I00) of the Recovery, Transformation and Resilience Plan. CIAEST/2022/1, Research Group in Public Law and ICT, Catholic University of Colombia; CIAEST/2022/1, Digital Rights Agreement-SEDIA Area 5 (2023/C046/00228673) and Area 6 (2023/C046/00229475).

[2] It should be noted that European Commission, Renda. A. (project leader), *Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe. Final Report (D5),* April 2021. https://op.europa.eu/es/publication-detail/-/publication/55538b70-a638-11eb-9585-01aa75ed71a1

On pp. 112 et seq. it can be seen that in the consultation process leading up to the proposed regulation, the definition of "high risk" was the most crucial point. Eighteen percent of the respondents (74 out of 408), considered this definition of "high risk" to be unclear or in need of significant improvement. The binary classification of risks as high or low had been perceived as oversimplified, leading several stakeholders to propose the introduction of additional levels of risk. Some argued that the current definition was too broad, while others believed it was too restrictive. Among the alternative proposals, at least six position papers had advocated the GDEC's stepwise approach, introducing five levels of risk for a more nuanced analysis. Others suggested the adoption of risk matrices, which combined the intensity of potential harm with the level of human involvement or control in Artificial Intelligence (AI) decisions. The likelihood of harm had also been repeatedly mentioned as an essential criterion to consider. In addition, numerous position papers criticized the proposed two-step approach ered in the EU White Paper on AI in 2020 and in the European Parliament's proposal of the same year have been significantly achieved . The Commission has stated – I do not know if with much foundation – that one third of public AI systems will be HRS,[3] while only 10% of private ones will be HRS.[4]

The basic idea of the AIA is that common standards for HRSs are required to ensure health, safety and fundamental rights (Whereas 7). This will ensure that output information from these systems used in the Union does not pose unacceptable risks to these public interests (Whereas 46).[5]

Obviously, it is recognized that this compliance system is a significant burden, so that "The classification of an AI system as an HRS should be limited to those AI systems which have a significant adverse effect on the health, safety and fundamental rights of individuals in the Union, and such limitation minimizes any possible restriction on international trade" (Whereas 46). Whereas 48 goes so far as to mention 21 fundamental rights that may be affected by HRS, in addition to the specific rights of minors.[6]

As will be explained, Article 6 contains the "Rules for the classification of high-risk AI systems" and follows a dual system. On the one hand, the consideration of HRS is related to products covered by certain listed Union har-

to determining "high-risk" AI. At least 19 of these papers considered it inadequate, and at least five opposed a sectoral approach. Other suggestions and criticisms varied widely. One notable proposal for improving risk assessment was to consider all subjects affected by the application of AI, stressing the importance of taking into account both collective and individual risks, as AI applications could entail risks to society as a whole, including democracy, the environment and human rights. The need to clarify the definition of "high risk" had been a concern shared by all stakeholders.

[3] JRC, Tangi, L. et al.: *AI Watch European landscape on the use of Artificial Intelligence by the Public Sector*, JRC Science for Policy Report, European Union. 2022, p. 58.

[4] European Commission, *Study to Support...* cit. p. 143.

[5] As stated in Recital 7 "It is appropriate to lay down common rules for high-risk AI systems in order to ensure a high and consistent level of protection of public interests as regards health, safety and fundamental rights". Therefore, 'The placing on the Union market, putting into service or use of high-risk AI systems should be subject to compliance by them with certain mandatory requirements, which should ensure that high-risk AI systems available in the Union or whose output information is used in the Union do not pose unacceptable risks to important public interests of the EU, recognized and protected by Union law. (Recital 46).

[6] These include: dignity, private and family life, data protection, freedom of expression and information, freedom of assembly and association, non-discrimination, the right to education, consumer protection, workers' rights, the rights of disabled persons, equality between men and women, intellectual property rights, the right to an effective remedy and to a fair trial, the rights of the defense and the presumption of innocence, and the right to good administration. In addition, the specific rights of minors and the health and safety of persons and the protection of the environment.

monization legislation or as safety components of these or as an independent product. On the other hand, and if general requirements are met, AI systems linked to purposes and uses listed in Annex III are considered as HRS.

Many of the AIA obligations relate to HRS and are mentioned in this work. I In any case, it is sufficient to recall that there is a specific adaptation period. In general, Annex III HRS obligations must be fulfilled 24 months after publication and Annex I HRS obligations, 36 months after publication (Art. 113). However, the public sector will have a privileged maximum period of six years (Art. 111.2). In any case, "Providers of high-risk AI systems are encouraged to start complying, on a voluntary basis, with the relevant obligations of this Regulation already during the transitional period." (Whereas 178).

II. A Warning: The Regulation of Systems as High-Risk by the AIA does not imply their legal entitlement

If being considered an HRS entails many consequences under the AIA, it is very important to remember that the regulation of an HRS by the AIA does not imply providing it with legal entitlement. As Whereas 63 of the AIA expressly warns, "The fact that an AI system is classified as a high-risk AI system under this Regulation should not be interpreted as indicating that its use is lawful under other acts of Union law or national law compatible with Union law [...] This Regulation should not be understood as constituting a legal basis [...] unless this Regulation specifically provides otherwise" (Whereas 63). This general rule is clear and I consider that it should apply to all HRS, being particularly relevant for Annex III HRS. This is despite the fact that, in a clear lack of legislative technique, only in three of the eight paragraphs of Annex III the expression "in so far as their use is permitted by the applicable Union or national law" is added. This is in the case for biometric AI systems (1st), for law enforcement (6th) and for the use for migration, asylum and border control management (7th). However, this need for specific regulation law is not added in the area of Justice (8th a).

In any case, it must be assumed that the AIA does not serve as a legal rule that legitimizes data processing, a restriction of fundamental rights or that fulfils a requirement of criminal, sanctioning or procedural legality. A law will continue to be necessary to enable the existence of a specific HRS of those regulated in general terms in the AIA.

On the other hand, and as an exception, the AIA expressly implies a regulation that provides a legal basis for legitimization in Article 10.5 regarding the possibility of using "exceptionally" special categories of data "to ensure the detection and correction of biases," under quite precise requirements.

III. Artificial Intelligence Systems in dangerous products of Annex I

The first set of HRS concerns AI systems in products that present certain levels of hazard and are therefore subject to the EU conformity assessment regime, which must be assessed by third parties. There is a certain complexity to the issue. These are products associated with certain EU harmonization legislation listed in Annex I (formerly Annex II in the process of drafting the regulation) or safety components of these products. According to Article 6.1, "an AI system shall be considered as high risk when both conditions are met", i.e. "it is intended to be used as a safety component [...] or *the AI system itself* is such a product" (a) and (b) "it must undergo a conformity assessment carried out by an independent body for its placing on the market or putting into service."

1. Artificial Intelligence system as a safety component or product in products subject to a third party "Conformity Assessment"

The aim of the AIA is to focus on AI applications that "may have an adverse effect on the health and safety of individuals" (Whereas 46). Thus, the concept of "safety component", already consolidated in the legal framework for machines, has been applied.[7] However, the idea of "safety component" is adjusted in the AIA to make it more general and now refers to "digital components"[8] which can be anticipated in respect of AI in all areas, as they fulfil a safety function for the product or system, or whose failure or malfunction endangers the health and safety of individuals or property (Article 3.14 of the AIA).[9]

[7] Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery. The definition in Article 3.3 can be followed. 3rd "'safety component' means a physical or digital component, including software, of a product falling within the scope of this Regulation which is designed or intended to perform a safety function and which is separately placed on the market, the failure or malfunction of which endangers the safety of persons, but which is not necessary for that product to function or for which normal components can be substituted in order for that product to function."

[8] Thus, Whereas 47 speaks of "safety risks that may be generated by a product as a whole due to its digital components, which may include AI systems".

[9] In particular, according to Article 3. 14th RAI "safety component" means a part of a

Thus, the AI system may or may not have been introduced on the market integrated into a product as a digital component of the product, either as a safety component itself or as a component that can generate danger, mentioning cases of robots or the health sector (Whereas 46). This includes cases of AI systems such as software or computer programs that are marketed separately from the product. AIA would also apply to independent software, which is considered a product in its own right. The most obvious example would be independent software for medical devices, regulated by Regulation 745/2017 on medical devices. A stand-alone AI system will be HRS according to "the severity of the potential harm as well as the likelihood of its occurrence," and the Commission will take them into account in its update of the list of HRS according to technological progress (Whereas 52). It should also be noted that in a number of sectors there is a preference for AIA to act indirectly, forcing them to adapt their specific regulations. This is the case where AI is incorporated as a safety component of products or systems in civil aviation, agricultural or forestry vehicles, marine equipment, railway systems, motor vehicles and their trailers. In these cases, the AIA obligations for HRS will have to be taken into account when adopting relevant delegated or implementing acts, tailored to the specificities of these sectors. The idea is not to "interfere with existing governance, conformity assessment and enforcement mechanisms and authorities" (Whereas 49). Thus, to illustrate what this is about, vehicle type-approval is governed by EU Regulation 2018/858, as amended by the AIA (Art. 107), so that the AIA has to be introduced in the delegated acts of this regulation.[1]

Another basic element for understanding this group of AI systems in Annex I is that it is subject to a third-party conformity assessment. At this point, it should be recalled that, for certain products, their placing on the market or putting into service can only take place when the product complies with all applicable Union harmonization legislation. This is in the context of the "New Legislative Framework" (Whereas 46). Annex I lists all these products.

In order for the AI systems of products subject to this harmonized legislation or safety components of such products listed in Annex I to be considered HRS, the harmonized legislation for such products must provide for

product or system which performs a safety function for that product or system, or the failure or malfunction of which endangers the health and safety of persons or property; If the definition is compared with that of the Machinery Regulation, the safety function becomes here an alternative and not the essential defining element.

[1] On the subject, VDA, *Position. Artificial Intelligence Act*, German Association on the Automotive Industry, Berlin, July, 2023, https://www.vda.de/en/news/publications/publication/artificial-intelligence-act

conformity assessment to be carried out by an "independent conformity assessment body in accordance with such Union harmonization legislation" (Whereas 50). It should be recalled that in the various pieces of legislation and within each one, depending on the type of product being regulated, in some cases this conformity assessment is a self-assessment, i.e. it is based on an internal control by the provider itself.[2] However, for products that are considered to be more hazardous, an independent third party must be involved.

Thus, for an AI system to be considered as an HRS, it is not sufficient that the AI system is a product or a safety component of a product in the Annex I harmonization legislation; in addition, that legislation must provide for the conformity assessment of the product to be carried out by a third party. This is justified because third party conformity assessment is considered as an indication that the product in question may have a negative impact on the safety or health of individuals and should therefore, for the purpose of AIA, be considered as an HRS. Therefore, there may be AI systems that are products or safety components of products covered by the harmonization legislation indicated by Annex I, but these are not considered as HRS because that legislation does not provide for third party conformity assessment. As recalled in Whereas 50: "Such products are, in particular, machines, toys, lifts, equipment and protective systems intended for use in potentially explosive atmospheres, radio equipment, pressure equipment, recreational craft equipment, cableway installations, equipment burning gaseous fuels, medical devices and in vitro diagnostic medical devices."

Similarly, an AI system may be an HRS for the AIA, but the product of which it is a safety component or product itself may not be an HRS in the particular regulatory domain that applies to that product (Whereas 51). This is the case for medical devices "which provide for an independent body to carry out a conformity assessment of medium and high-risk products" (Whereas 51).[3] Therefore, if they are not assessed as HRS by that body, they would not be HRS for the purposes of the AIA.

If the system is itself neither a product nor a safety component, but an

[2] Please refer to the specific section in this work by Adrián Palma. For more information on conformity assessments https://single-market-economy.ec.europa.eu/single-market/ce-marking/manufacturers_en

[3] 51 The fact that an AI system is classified as high risk under this Regulation does not necessarily mean that the product of which it is a safety component, or the AI system itself as a product, is considered to be 'high risk' under the criteria laid down in the relevant Union harmonization legislation applying to the product. This is the case, in particular, of Regulations (IA) 2017/745 and (IA) 2017/746, which provide for an independent body to carry out a conformity assessment of medium and high-risk products.

"independent AI system," its purpose and risk will have to be considered. It will be up to the Commission to determine whether it is considered an HRS through delegated acts (Whereas 52).[4] It should be noted that the issue becomes extraordinarily complex. As Palma analyses in this work, it will be necessary to follow each Regulation that controls these products in a particular way, with special attention to Regulation 2023/1230 on machinery or the aforementioned Regulations 745/2017 and 746/2017 on medical devices. Moreover, these regulations will have to be adapted to the subsequent approval of the AIA, which they obviously do not currently take into account.

2. The provider must reasonably know whether its Product can be in Annex I

Whoever develops an AI system specifically for incorporation into products or as safety components should be aware of their industry, the usual nature of the users of the AI system and the legal regime that applies to these products. In other words, it is natural for the producer of such products or the AI service provider to know whether the system is subject to the specific Union harmonization rules referred to in Annex I. On this basis, it will be able to assess whether the conditions of Article 6 are met for it to be considered as an HRS. Thus, it shall take into account whether the harmonized legislation of the product or safety component of the product provides for conformity assessment by third parties. If so, the service provider will have to develop the AI system as high risk by fulfilling all requirements.

The provider of an AI system that is not specifically developed for incorporation into such products must also assess whether its AI system alone can reasonably and potentially be considered a product covered by Annex I harmonization legislation. In terms of the AIA , it must assess what the "intended purpose" is (art. 3. 1º. 12º).[5] If so, in similar terms to what happens

[4] "As regards stand-alone AI systems, i.e. those high-risk AI systems which are not safety components or which are not products in themselves, they should be classified as high risk if, in the light of their intended purpose, they present a high risk of being harmful to the health and safety or fundamental rights of individuals, taking into account both the severity of the potential harm and the likelihood of its occurrence, and are used in a number of pre-defined areas specified in this Regulation. For the identification of such systems, the same methodology and criteria are used as foreseen for the possible future modification of the list of high-risk AI systems, which the Commission should be empowered to adopt, by means of delegated acts, in order to take into account, the rapid pace of technological development, as well as possible changes in the use of AI systems."

[5] "(12) 'intended purpose' means the use for which an AI system is intended by a provider, including the specific context and conditions of use, as indicated by the information

with respect to the possibility of using a system for Annex III purposes, you need to know whether the specific legislation requires a third-party conformity assessment or specifically regulates it as an HRS. If so, it will be an HRS with all the consequences that this entails.

A provider should also assess whether its AI system can reasonably and potentially be integrated into a product or used as a safety component of a product subject to the harmonized Annex I legislation. If it can reasonably foresee that this is the case, and also because of its go-to-market strategy, the provider should consider its AI system to be an HRS for the purposes of regulatory compliance with AIA obligations. This will be necessary in order to be able to bring it to the market or make it available as such. In any case, for potential customers or users of the system you develop, you will need to design a legal framework determining whether or not the AI system can be incorporated into products or as a security component. In this way, your entity will have complied with the AIA's requirements and the user will have to comply with yours.

IV. High-Risk Systems pursuing the purposes of Annex III

1. Systems having a substantial influence on decision-making for the Purposes of Annex III and the criteria for considering it

In principle, "the AI systems referred to in Annex III shall be considered as high risk" (Art. 6. 2. AIA). Annex III on "High-risk AI systems referred to in Article 6(2)" groups the types of HRS under eight headings: 1. Biometrics; 2. Critical infrastructure; 3. Education and vocational training; 4. Employment, management of workers and access to self-employment; 5. Access to and enjoyment of essential private services and essential public services and benefits; 6. Law enforcement; 7. Migration, asylum and border control management; 8. Administration of justice and democratic processes. These eight sections of Annex III contain twenty-five letters with as many types of AI systems by purpose. Twenty-five times the formula "AI systems intended for use..." is used.[6]

provided by the provider in instructions for use, sales and marketing materials and statements, and technical documentation;'.

[6] Strictly speaking, on 24 occasions the formula is used to describe the purpose defining high risk. However, the first occasion (Annex III 1. a) is a set "1. Biometrics, insofar as their use is permitted by applicable Union or national law". This set includes three cases. Points b and c do follow the wording "AI systems intended to be used", however, as an exception to the

Thus, the "intended purpose" (art. 3. 12 AIA)[7] for which the AI system is designed and "the capability of an AI system to achieve its intended purpose," i.e. its "functioning" (art. 3. 18 AIA),[8] is the determining element of Annex III. Therefore, and as will be specified, if the AI system developed has as its intended purpose one of those in Annex III and the capability to achieve it, it must be presumed to be an HRS and only exceptionally will it not be an HRS.

However, it is important to note how the AIA has changed in its wording on this point up to its final version. Initially, there was an automatism: the AI system was an HRS if it was intended for the purposes described in Annex III. Later, in the December 2022 EU Council version, the exception was added that it would be an HRS "unless the information output from the system is *merely incidental* to the relevant action or decision to be taken" (Art. 6.3 AIA December 2022 EU Council). However, in the final version, there is no longer such automatism, but in addition to the purposes of Annex III, certain requirements must be met for it to be considered high risk: the AI system with the purposes of Annex III must effectively generate a risk and, above all, "substantially influence the outcome of decision-making" (Art. 6. 3º AIA).

If I may say so, this is a *substantial* change whereby the AI system has to *substantially* influence the decision to be taken. Indeed, the seasoned reader may be *automatically* led to think of the regulation of *automated* decisions in Article 22 GDPR[9] or Article 9.1 of Council of Europe Convention 108 in its 2018 version. It should be recalled that the special guarantees provided

whole Annex, in point a) the wording "AI systems intended to be used" is used as an exception for remote biometric identification systems, i.e. "AI systems intended to be used for biometric verification purposes whose sole purpose is to confirm that a specific natural person is the person he claims to be shall be excluded".

[7] "(12) 'intended purpose' means the use for which an AI system is intended by a provider, including the specific context and conditions of use, as indicated by the information provided by the provider in the instructions for use, sales and marketing materials and statements, and technical documentation;'.

[8] "(18) "performance of an AI system" means the ability of an AI system to achieve its intended purpose;'.

[9] I have analyzed this precept in particular in "Derechos y garantías ante el uso público y privado de inteligencia artificial, robótica y big data", in Bauzá, M. (dir.), *El Derecho de las TIC en Iberoamérica*, Obra Colectiva de FIADI (Federación Iberoamericana de Asociaciones de Derecho e Informática), La Ley – Thompson-Reuters, Montevideo, 2019, pp. 917-952, accessed at http://links.uv.es/BmO8AU7. In any case, for all, Palma Ortigosa, A., *Decisiones automatizadas y protección de datos personales. Especial atención a los sistemas de inteligencia artificial*, Dykinson, 2022 and Roig I Batalla, A., *Las garantías frente a las decisiones automatizadas del Reglamento general de Protección de Datos a la gobernanza algorítmica*, J.M. Bosch, Barcelona, 2021.

by Article 22 GDPR are in respect of "a decision based *solely* on automated processing," whereas it will be HRS of Annex III whenever it substantially influences the decision to be taken.

In the field of data protection, the issue has been important since, in principle, decisions that are not solely automated do not enjoy the guarantees of Article 22 GDPR. However, the interpretation by authorities and judges has moved in a clearly pro-guarantee direction to protect also decisions that are apparently human but substantially based on the automated system. For the Article 29 Working Party (EDPB), it does fall within the scope of Article 22 GDPR "if someone routinely applies automatically generated profiles to individuals without this [human review] having any real influence on the outcome, this would still be a decision based solely on automated processing." For this right not to apply, the human intervention must be "meaningful, rather than merely a token gesture" and carried out by "an authorized and competent individual."[10]

Particularly noteworthy is the judgment of the CJEU of 7 December 2023,[11] the first to deal centrally with Article 22 GDPR. It entails a "piercing of the veil" of "solely" automated decisions. Thus, the guarantees of this provision will also apply if the results of the automated system, profiling or automated weighting of data (or with Artificial Intelligence) are materially connected with the decision finally incorporated by the person who has to adopt it with respect to the person affected by said decision, despite the fact that there may be human mediation or intervention. In this line, some steps were already being taken by the authorities.[12] The most recent rules are also recognizing guarantees for partially or semi-automated decisions.[13] It should

[10] G29-EU, *Guidelines on Decisions* ... cit. p. 23.

[11] The first study on the same Cotino Hueso, L. "La primera sentencia del Tribunal de Justicia de la Unión Europea sobre decisiones automatizadas y sus implicaciones para la protección de datos y el Reglamento de inteligencia artificial", *Diario La Ley*, January 2024. https://ir.uv.es/V14YNLl Access to the judgement at https://curia.europa.eu/juris/document/document.jsf?text=&docid=280426&pageIndex=0&doclang=ES&mode=lst&dir=&occ=first&part=1&cid=10472490

https://eur-lex.europa.eu/legal-content/es/TXT/?uri=CELEX:62021CJ0634

[12] The Portuguese data protection authority considered as fully automated a process that included human intervention, as the person in charge of monitoring the results of the algorithm had no defined guidelines or criteria for its interpretation. *Comissão Nacional de Proteção de Dados. Deliberação* 622/2021. Paragraph 55. Resolution available at: https://www.cnpd.pt/decisoes/deliberacoes/.

[13] Thus, Article 20 of Ecuador's Organic Law on the Protection of Personal Data (Registro Oficial Suplemento 459 of 26-May-2021) extends the right to "a decision based solely or partially on automated assessments". In Canada it is worth noting the *Directive on Automat-*

be borne in mind that automated – or AI-enabled – decisions are often part of a chain or ecosystem of actions in which there is human intervention. Therefore, the existence of such human intervention should not *automatically* exclude the application of the particular guarantees conferred by Article 22 for automated decisions. However, and as far as we are concerned here, the AIA leaves behind the need for the decision to be solely automated and expressly defines a series of criteria for considering whether the AI system substantially influences the decision.

About the Criteria for Considering that the Artificial Intelligence System has a Substantial Influence on the Decision.

The premise is that the system "poses a significant risk of harm to the health, safety or fundamental rights of natural persons," but that risk must be "in particular" by "substantially influencing the outcome of decision-making" (Art. 6.3 AIA). Thus, "an AI system that does not affect the substance, and therefore the outcome, of human or automated decision-making" (Whereas 53) is not considered to be an HRS. As detailed below, systems used for the purposes of Annex III, but with a "limited procedural" task, or to "enhance [...] a previously performed human activity;" where the system is "not intended to replace [...] or influence" decision-making, but to "detect patterns" or, finally, where the use of the AI system is clearly ancillary, will not be HRS.

Without prejudice to the additions in this respect by the EU Council in December 2022 and especially by the Parliament in June 2023, it has been specified in the version finally agreed that this substantial influence will not occur "where one or more of the following conditions are met:"

a) "The AI system is intended to perform a limited procedural task". In this respect, it is specified that it would not be HRS "an AI system that transforms unstructured data into structured data, an AI system that categorizes incoming documents or an AI system that is used to detect duplicates among a large number of applications. The nature of these tasks is so restricted and limited that they present only limited risks" (Whereas 52). It will have to be determined on a case-by-case basis when the task is a "limited procedure."

b) "the AI system is intended to improve the outcome of a previously performed human activity". On this point, whereas 52 specifies that "the AI

ed Decision-Making which since its first version in 2018 which defines an automated decision system as "Any technology that assists or replaces the judgment of human decision-makers." (Appendix A – Definitions), so by no means are the guarantees of this Directive limited to fully automated decisions.

In the case of Spain, for example, the recent Charter on Digital Rights contemplates the need to carry out an impact assessment on digital rights when algorithms are designed for automated or semi-automated decision-making (section XVIII.7).

system only adds an additional layer to the human activity, thus entailing a lower risk. This condition would apply, for example, to AI systems intended to improve the language used in documents already drafted, for example, as regards the use of a professional tone or an academic linguistic register or the adaptation of the text to a particular brand communication." In this case, the action and the basic elements of the decision to be taken must be human and prior to the use of AI. The evidentiary element will be essential.

c) "The AI system is intended to detect patterns of decision making or deviations from previous patterns of decision making and is not intended to replace or influence the previously made human assessment without appropriate human review." With respect to this assumption, it is specified that "the risk would be lower because the AI system is used following a previously made human assessment and is not intended to replace or influence it without appropriate human review. For example, AI systems of this type include those that can be used to check *a posteriori* whether a teacher may have deviated from his or her particular grading pattern, in order to draw attention to possible inconsistencies or anomalies" (Whereas 52). In this case, the substantive element seems to be that the AI system is not intended to make or influence decisions, but to evaluate decisions made by humans.

d) "The AI system is intended to carry out a preparatory task for an assessment relevant to the use cases listed in Annex III." Whereas 52 is again interesting, as it specifies that "the potential impact of the outcome of the system would be very low in terms of posing a risk for the subsequent assessment. This condition covers, inter alia, intelligent solutions for records management, including various functions such as indexing, searching, text and speech processing or linking data to other data sources, or AI systems used for the translation of the initial documents. In this case, the clearly ancillary and remote nature of the decision to be taken seems to be the distinguishing feature.

Thus, it would be sufficient if only one of these circumstances is considered to be present for the system to be considered as a non-Annex III HRS. Obviously, if several of these circumstances are present, it will be clearer that it is not an RAS. It should be recalled that an AI system used for Annex III purposes is presumed to be an Annex III HRS, and that the exception is that it is not, by proving and justifying that one of these circumstances is present. Similarly, Article 7 AIA, discussed below, sets out a number of criteria that can be used for interpretation and application in each specific case.

2. Profiling with Artificial Intelligence for the purposes of Annex III shall always be High-Risk

As a premise, "AI systems referred to in Annex III shall always be considered high-risk where the AI system carries out profiling of natural persons""(Art. 6.3). According to Whereas 53, it is appropriate to refer to the GDPR definition (Art. 4.4) of 'profiling.'[14] Thus, an AI system that, for the purposes of Annex III, assesses[15], analyses or predicts aspects relating to professional performance, economic situation, health, personal preferences, interests, reliability, behavior, location or movements on the basis of data relating to a natural person will always be an HRS. However, it would not be sufficient for the AI system to perform 'simple classification' if it is only 'to obtain an overview of these [persons] without making predictions or drawing conclusions about an individual.[16] Following the ECDC, it would be a matter of using an automated system that integrates AI to process personal data to make such an assessment or analysis, always in relation to the purposes of Annex III. It is important to note that profiling with AI that is partial and not total would be HRS, as "any form of automated processing" is valid.

If these requirements are met, such AI profiling for the purposes of Annex III will be HRS and, as a consequence, the obligations of the AIA will apply. In addition, of course, any relevant GDPR regime will also apply to such profiling (Whereas10 GDPR)[17] . Among the data protection rules, the particularities of fully automated decisions Article 22 GDPR will often – but not always – have to be applied.[18]

[14] The definition is the same in Article 4(4). Directive (EU) 2016/680 of 27 April 2016: '(4) "profiling" means any form of automated processing of personal data consisting in using personal data to evaluate certain personal aspects relating to a natural person, in particular to analyze or predict aspects concerning that natural person's professional performance, financial situation, health, personal preferences, interests, reliability, behavior, location or movements;'.

[15] Recalls the ECDC that "The use of the word "assess" suggests that profiling involves some kind of evaluation or judgement of a person." Article 29 Working Party, *Guidelines on automated individual decisions and profiling for the purposes of Regulation 2016/679*, 3 October 2017, final version 6 February 2018, Doc WP251rev.01, https://www.aepd.es/documento/wp251rev01-en.pdf, p. 6-7.

[16] *Idem.*

[17] Whereas 10. [...] It should also be made clear that data subjects continue to enjoy all the rights and guarantees conferred on them by Union law, including rights related to fully automated individual decisions, such as profiling.

[18] In this respect, G29-EU, *Guidelines on decisions* ... cit. p. 8-9 recalls that automated decisions have a different scope of application and may overlap with or partially derive from profiling. It is recalled that automated decisions can be carried out with or without profiling;

Finally, with regard to profiling with AI it should be taken into account that they may fall under the specific prohibitions (Art. 5.1. d) AIA, Whereas 42).[19] Also, the specialties of the AI systems for the application of the law must be taken into account (Annex III 6. d and e).[20]

3. The purposes of High-Risk systems in Annex III

Many of the cases that are considered to be HRS under Annex III are discussed in more detail elsewhere in this book, and reference can be made to them in general. However, it is worth recalling the main purposes and scenarios of Annex III HRS.

3.1. Not Banned Biometrics , Infrastructure, Education and Work

Firstly, "biometric" AI systems are HRS. It should be recalled that they will be HRS as long as they are not *totally banned* by Article 5. And it is stated that they *are totally* insofar as, especially in the case of "real-time" remote biometric identification in public access areas for law enforcement purposes (Art. 5. 1º h), the possible legal exceptions and under an authorization system

profiling can take place without carrying out automated decisions. However, the two are not necessarily independent activities. Something that starts as a simple automated decision process can develop into a process based on profiling, depending on the use made of the data. In this direction he points out as an example the imposition of speeding fines solely on the basis of speed camera evidence is an automated decision process that does not necessarily involve profiling. It is noted that it would be a profiling-based decision if the person's driving habits are monitored over time and, for example, the amount of the fine imposed is the result of an assessment involving other factors, such as whether the speeding is a repeat offence or whether the driver has recently committed other traffic offences. It is also noted that decisions that are not based solely on automated processing may also involve profiling. For example, before granting a mortgage, a bank may take into account the borrower's credit rating, and additional significant human interventions may take place before any decision on the individual is made.

[19] These would be "risk assessments [...] for the purpose of assessing or predicting the likelihood of a natural person committing a criminal offence on the basis of profiling alone". And Recital 42 states that "natural persons should never be judged on the basis of behavior predicted by an AI based on profiling alone".

[20] In principle, it would not be high risk to profile data with AI. Annex III 6. D) indicates that the use of AI "to assess the likelihood of a natural person to commit an offence or to commit a repeat offence not only for the purpose of profiling natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 or to assess personality traits and characteristics or past criminal behavior of natural persons or groups" is high risk.

On the other hand, according to subparagraph (d), AI profiling 'during the detection, investigation or prosecution of criminal offences' is high risk. And under point (e), AI systems 'for profiling of natural persons [...] during the detection, investigation or prosecution of criminal offences' would be high risk.

are foreseen. Thus, if the system is not banned, it will in any case be HRS. This is also the case for cases that are not banned under the general ban of systems for "inferring emotions" (Art. 5. 1º f). Similarly, "biometric categorization systems that individually classify people" but not for the sensitive purposes of Article 5.1.g. would be HRS because they are not banned .

Secondly, under "biometrics" in Annex III, it is specified that "remote biometric identification systems" (a) are HRS. In particular, it should be noted that one-to-one biometric identification systems, i.e. "the sole purpose of which is to confirm that a specific natural person is the person he or she claims to be," are not HRS. This is relevant because, although the ECDC[21] and the AEPD[22] recently considered that the processing of data for this purpose of one-to-one identification is indeed a processing of sensitive data category of Article 9 GDPR, it would not be a HRS for the purposes of the AIA because of the express exclusion in Annex III. Biometrics would also include systems "for biometric categorization on the basis of sensitive attributes or characteristics" or "for the recognition of emotions," in the terms specified in this paragraph 1.

As regards AI systems in "critical infrastructures,"[23] there is a certain proximity to Annex I in that it deals with the concept of "security components" (2.a), for which reference should be made to what has already been said. In any case, it is specified that "these are systems used to directly protect the physical integrity of critical infrastructure or the health and safety of individuals and property, but which are not necessary for the operation of the system." Examples include "water pressure monitoring systems or fire alarm control systems in cloud computing centers." And it is relevant that "components intended to be used exclusively for cybersecurity purposes should not be considered as security components" (Whereas 55).

With regard to critical infrastructures, these are generally defined in Article 3. 62nd AIA, which in turn refers to Article 2, item 4, considering "critical digital infrastructures"[24] as those "listed in Annex I, item 8, of Directive (EU)

[21] Thus, the ECDC in May 2022 initiated this change in the first version of the ECDC, *Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement* and on 26 April 2023 updated and reinforced this criterion in Version 2.0 https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052022-use-facial-recognition-technology-area_en.

[22] The AEPD has externalized this change of criteria in AEPD, *Guía Tratamientos de control de presencia mediante sistemas biétricos* de 23 de noviembre de 2023. https://www.aepd.es/guias/guia-control-presencia-biometrico.pdf

[23] In the regulatory process both the European Commission in 2021 and the European Parliament in 2023 spoke of the "Management and Operation" of infrastructure.

[24] In the RAI pipeline, the mention of critical digital infrastructures appears in the EU Council's version in December 2022.

2022/2557" (Whereas 55). Critical infrastructures are also those relating to "road traffic or the supply of water, gas, heat or electricity" (2.a). In principle, only "road traffic" would be included and therefore not air or rail traffic.[25] The explanation is obvious: "a failure or malfunction of these components can endanger human life and health on a large scale and significantly disrupt the normal course of social and economic activities" (Whereas 55). In any case, account should be taken of specific legislation and the possibility that an Annex II HRS is involved.

In the field of education (Annex III. 3º), it is assumed that "promoting high quality digital education and training" (Whereas 56) can be positive. However, some specific uses are considered to be HRS because they "may particularly encroach on and violate the right to education and training, and the right to non-discrimination, as well as perpetuate historical patterns of discrimination" (Whereas 56). Throughout the regulatory process, these uses were delimited and expanded.[26] Thus, systems that determine "access or admission" or "to allocate individuals" between educational institutions are specified as HRS (3.a). Also, AI systems for the assessment of learning outcomes, in this sense it is underlined that "in particular when these outcomes are used to guide the learning process of individuals" (3.b). Likewise, following the EU Parliament's proposal, it is also HRS AI "to assess the appropriate level of education that a person will receive or be able to access" (3.c). Finally, also proposed by the Parliament, systems "for the monitoring and detection of prohibited behavior [...] during examinations" are considered to be HRS.

The field of employment has been gaining presence throughout the AIA approval process. Finally, AI systems for "Employment, management of workers and access to self-employment" (Annex III. 4º) will be HRS because they "may significantly affect future employment prospects, the livelihoods of such individuals and the rights of workers" and "may perpetuate historical patterns of discrimination [...] throughout the recruitment process and in the assessment, promotion or retention of persons in contractual relationships of an employment nature" (Whereas 57).

More specifically, they will be HRS if they are used in particular "for the recruitment or selection of natural persons, in particular for publishing specific job advertisements, analyzing and filtering job applications and as-

[25] It should be noted that the final version refers to "road traffic", the Council version of December 2022 mentions only "road traffic" and the Parliament added "rail and air", but this was not adopted in the final version.

[26] This can be seen from the EU Council's version, refined by the Parliament, which is essentially the one finally adopted.

sessing candidates" (4.a).[27] Also for making "decisions affecting the terms of employment relationships or the promotion or termination of employment relationships, for the allocation of tasks on the basis of individual behavior or personal traits or characteristics or for monitoring and evaluating performance and behavior" (4.b).

3.2. Essential Services and Benefits: Administration, Emergencies, Insurance and Banking

Another group of HRS are those relating to "access" and "enjoyment" of "essential private services and essential public services and benefits" (Annex III. 5). The consideration of "essential" is due to the addition of the EU Council in its December 2022 version. It is striking that public and private use of AI systems are addressed together. This is justified on the grounds that they are "essential services and benefits [...] necessary to enable people to participate fully in society or to improve their standard of living, and the enjoyment of such services and benefits" (Whereas 58).

Regarding the use of AI by the Administration and the public sector, reference should be made to the corresponding section in this work. I would just like to point out now that, although they are all there are, they are not all there are; that is, there is no doubt that the public use of biometric, law enforcement, education, labor or critical infrastructure systems should be considered HRS.[28] However, it seems that the use of AI systems by the administration that impacts on rights only focuses on "essential public assistance benefits and services".[29] These include a very broad spectrum of administrations ("health", "social security," "social services" related to "maternity," "accidents at work, dependency or old age and loss of employment, social assistance and housing support"). It should be recalled that the use of AI systems by the public sector enjoys a – disproportionate – 6-year compliance period.

However, many public uses of AI that also impact on more than a few fundamental rights seem to be left out of the high risk. Among others, I would now like to highlight the increasingly common systems for the prosecu-

[27] In the regulatory process, the purpose of "advertising vacancies" has been omitted in favor of "publishing advertisements".

[28] In this respect, among other studies, I refer to my study "Los usos de la IA en el sector público, su variable impacto y categorización jurídica" *Revista Canaria de Administración Pública*, n.º 1, 2023, pp. 211-242, accessed in journal, accessed in article.

[29] In the Commission's initial version of 2021 or the Council's version of December 2022, there was no reference to any specific area. The Parliament mentioned "health services and essential services, inter alia housing, electricity, heating/cooling and internet", but in the end it was decided not to specify these services in the public sector in this letter.

tion of fraud or taxation,[30] which could be considered for "law enforcement" (Annex III. 6), but which in many cases fall outside the criminal proceedings of the said paragraph 6.[31] This intention is expressly stated in Whereas 59, which expressly excludes some systems used in "administrative processes by tax and customs authorities and financial intelligence units."[32] May I also draw attention to the fact that, had Parliament's Amendment 738 been followed, systems used by any administrative body for "the investigation and interpretation of facts and law, as well as in the application of the law to a particular set of facts" would have become HRS , as is the case for the use of AI systems in justice. This would have made a big difference to the projection of AIA to the public sector, but this is not the case.

As regards the scope of *emergencies*, paragraph 5.d includes AI systems to be used by the public sector, such as those relating to "emergency calls [...] priorities in [...] emergency situations [...] police, fire and medical assistance services, and in patient triage systems." The Commission's initial version spoke of 'emergency, including fire and medical aid', and the final clarifications were subsequently introduced by the Council and the Parliament.

More closely linked to the private sector, this group of the 5th essential services includes creditworthiness (5.b). Previous versions excluded AI systems "operated by small-scale providers for their own use" or "by providers which are micro and small enterprises." However, they are not finally excluded, without prejudice to the application of Article 63. However, since the Parliament's version in June 2023, those used for the detection of financial

[30] In this respect I refer to my study "Hacia la transparencia 4.0: el uso de la inteligencia artificial y big data para la lucha contra el fraude y la corrupción y las (muchas) exigencias constitucionales", in Ramió, C. (coord.), *Repensando la administración digital y la innovación pública*, Instituto Nacional de Administración Pública (INAP), Madrid, 2021. https://links.uv.es/FU-W2pz6 For the tax field, Olivares, B. D., "Law and Artificial Intelligence in the Spanish Tax Administration: the Need for a Specific Regulation", *European Review of Digital Administration* & Law-ERDAL 1 (1-2), pp. 227-234.

[31] It should be noted that "AI systems used for the purpose of detecting financial fraud" are expressly excluded as high risk (5.b). However, Whereas 58 does not seem to refer to public fraud detection systems. It would, however, be "AI systems provided for by Union law with a view to detecting fraud in the provision of financial services and, for prudential purposes, for calculating the capital requirements of credit institutions and insurance undertakings should not be considered as high risk under this Regulation".

[32] In particular, 'AI systems specifically intended for use in administrative processes by tax and customs authorities and financial intelligence units carrying out administrative tasks of information analysis in accordance with Union anti-money laundering law should not be classified as high-risk AI systems used by law enforcement authorities for the purpose of preventing, detecting, investigating and prosecuting criminal offences'.

fraud are excluded. The consideration of creditworthiness services as HRS is justified as it implies possible access to essential resources and services such as "housing, electricity and telecommunication services" (Whereas 58).

Although not in the Commission's initial 2021 version, in the various versions of the EU Council, though not in its December 2022 proposal, AI schemes appeared and disappeared from the insurance field. From the Parliament in June 2023, systems for "the eligibility of natural persons for health and life insurance" were proposed as HRS. In the final adopted version, AI systems for "risk assessment and pricing in relation to natural persons for life and health insurance" are HRS (5.c). This is because they "may significantly affect people's livelihoods and [...] may violate their fundamental rights and have serious consequences for people's lives and health, such as financial exclusion and discrimination."

3.3. Law Enforcement, Migration, Asylum, and Border Control Management, Justice and Democratic Processes

As for law enforcement systems (Annex III. 6), undoubtedly one of the most important systems in Annex III, reference should also be made to the chapter on this issue in this work, recalling that they are focused on the area of criminal prosecution. By affinity, it is worth projecting what is stated there with regard to the AI HRS systems for "Migration, asylum and border control management" (Annex III. 7). In this area, it is reiterated that the possibility of using AI systems will only be possible if there is specific authorization and regulation.[33] In addition, special consideration should be given to the procedural requirements established by the applicable regulations.[34]

In the last section, AI systems for the "Administration of Justice and Democratic Processes" (Annex III. 8) are considered as HRS.[35] In this area, the scope has been further refined since the first version in 2021. In the final version, AI systems "intended for use by a judicial authority" (8.a) to "assist" in the investigation and interpretation of facts and law, as well as in the application of the law to a particular set of facts, are HRS. Also included are

[33] Recital 60: 'in so far as their use is permitted under Union and national law'.

[34] Such as Regulation (EC) No 810/2009 of the European Parliament and of the Council or Directive 2013/32/EU of the European Parliament and of the Council. Such as Regulation (EC) No 810/2009 of the European Parliament and of the Council or Directive 2013/32/EU of the European Parliament and of the Council.

[35] Thus, the Commission's 2021 version spoke of systems to "assist a judicial authority in the investigation and interpretation of facts and law and in the application of the law to a particular set of facts". To which was added that the systems may be "used by or on behalf of a judicial authority" (from EU Council version).

ADR systems "similarly used in alternative dispute resolution" (8.a), to which is added "where the results of alternative dispute resolution procedures produce legal effects for the parties" (Whereas 61).

The AIA excludes as HRS "AI systems intended for purely ancillary administrative activities that do not affect the actual administration of justice in specific cases, such as anonymization or pseudonymization of court decisions, documents or data, communication between staff members or administrative tasks" (Whereas 61).

The delimitation of substantive jurisdictional use with respect to administrative and non-jurisdictional use is an issue that is not always clear-cut and may generate future interpretative problems.[36] In fact, among other issues, it may determine the authority that will be competent for the supervision of AI systems for the Administration of Justice, be it the AEPD, the AESIA or the *CGPJ*.[37]

Finally, as proposed by the Parliament (Amendment 739), AI systems for "democratic processes" are included as HRS, in particular "to influence the outcome of an election or referendum or the voting behavior of natural persons exercising their right to vote in elections or referenda." It is noted that systems "to whose output information natural persons are not directly exposed, such as tools used to organize, optimize or structure political campaigns from an administrative or logistical point of view" (Annex III 8.b) and Whereas 62) will not be subject to HRS.[38] It should be recalled that Parliament's Amendment 740, which would have provided for HRS recommendation systems for large regulated platforms in the DSA, was ultimately unsuccessful.[39] On this issue, it is worth referring to the entire chapter of this work on the treatment of large platforms and Artificial Intelligence systems aimed at political influence, where the connection with other rules such as the DSA, the DMA and, in particular, with the recently approved Regulation

[36] It should be recalled that for the purposes of data protection and the competent authority, Article 236a LOPJ distinguishes between processing of personal data carried out for jurisdictional and non-jurisdictional purposes. The processing of data included in proceedings for the purpose of the exercise of judicial activity is considered to be jurisdictional.

[37] In this respect, my study "El uso jurisdiccional de la inteligencia artificial: habilitación legal, garantías necesarias y la supervisión por el CGPJ", *Actualidad Jurídica Iberoamericana*, n.º 21, 2024, monographic. https://revista-aji.com/

[38] Recital 62 makes no contribution in this case.

[39] "(ab) AI systems intended to be used by social networking platforms designated as very large online platforms within the meaning of Article 33 of Regulation (EU) 2022/2065, in their recommender systems to recommend user-generated content available on the platform to the recipient of the service."

(EU) 2024/900 of the European Parliament and of the Council of 13 March 2024 on transparency and targeting in political advertising, is especially taken into account.

4. Presumption that the Artificial Intelligence System pursuing Annex III purposes is indeed High Risk. Special obligations and actions

It is possible that a provider developing an AI system for Annex III purposes may consider that it is not an HRS. As noted above, if the intended purpose (art. 3.12 AIA) for which the AI system is designed coincides with Annex III purposes, it should be presumed to be an HRS and only exceptionally not an HRS.

For these cases, providers must act automatically and, in addition, there is a protocol for action by the market surveillance authority. Thus, in these cases, providers must document that their system is not an HRS and, in any case, they must include them in the register (Art. 6. 4º and Whereas 52).[40] This is a striking duty imposed on AI providers who consider their system not to be an HRS , but are presumed to be one. On the other hand, Article 80 states that, if the authority suspects that it might be an RAS, it must assess it. If the assessment reveals that the AI system is indeed HRS, the authority will require the provider to take the necessary measures to comply with the AIA and correct the problem within a timeframe set by the authority. If the use of the AI system goes beyond the national level, the supervisory authority must inform the European Commission and the other Member States about the assessment and the measures required of the provider.

I consider that these situations may arise especially in cases where the provider develops a system that *potentially* serves certain Annex III purposes, but *does not control the specific use that will be made by the user or deployer of the system*. In other words, the provider of an AI system for hiring, monitoring or dismissing employees is not the one who hires, monitors or dismisses the company's employees. In such cases, the provider must assess whether the

[40] Thus, Article 6(4): "A provider who considers that an AI system referred to in Annex III is not high risk shall document his assessment before such a system is placed on the market or put into service. On request of the competent national authorities, the provider shall provide the documentation of the assessment". Recital 52 states that "In order to ensure traceability and transparency, suppliers who, on the basis of these conditions, consider that an AI system is not high risk, should draw up the documentation of the assessment prior to placing on the market or putting into service of such a system and provide it to the competent national authorities upon request. Such providers should be obliged to register the system in the EU database set up under this Regulation.

inferences generated by the AI system can become a substantial element supporting the purposes of Annex III by the user of its system. If it reasonably believes this to be the case, the provider should assume that its system is HRS and, accordingly, develop it in accordance with the requirements of the AIA. In such cases, it is not sufficient for the provider to simply state in its instructions for use or the contractual framework with the user or deployer that it is not to be used as a substantive element in making decisions regarding the purposes of Annex III. The system will be an HRS if it generates outputs that make it suitable to inform decisions for these purposes. In any case, the specific case will have to be analyzed and also the possibility for the provider to invoke Article 6.4 of the AIA and to document and justify that the system is not an HRS. The paradoxical situation is that providers have obligations to market surveillance authorities, and these authorities will mostly act if they consider that their system is not an HRS . Whereas, if the system is HRS, the provider may well *only* have to do a self-assessment of compliance and may not be supervised by the authority.

5. When the deployer alters a system to a High-Risk Purpose

Another situation may arise where a provider develops an AI system that is not initially intended for Annex III purposes, but could be altered in its system or purpose to accommodate them. This would be considered a "reasonably foreseeable misuse" (Art. 3.13),[41] essentially by the user or deployer who intends it for Annex III purposes.

In this case, a distinction must be made between the unlikely event that such an AI system is already an HRS. In such a case, the provider has certain obligations. It must manage the risks and assess the likelihood of such misuse for an Annex III purpose, and provide for measures to mitigate these risks and their effects (Art. 9.2(b) AIA[42] and Whereas 65). In addition, if any risk exists, it must inform the user or deployer of the system in the instructions for use. This risk of use for Annex III purposes must also be prevented by human control (Art. 14. 2º AIA).[43] In any case, if the deployer makes a

[41] "“Reasonably foreseeable misuse’ means the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behavior or interaction with other systems, including other AI systems.

[42] Article 9. Risk management system [...] 2nd (b) "the estimation and evaluation of the risks that could arise when the high-risk AI system is used in accordance with its intended purpose and when it is put to reasonably foreseeable misuse".

[43] Article 14 Human oversight. [2nd "The objective of human oversight shall be to prevent or minimize risks to health, safety or fundamental rights that may arise when a high-risk

"substantial modification" to the HRS, e.g. to directly make decisions in the context of Annex III purposes, he will assume the obligations of the provider (Art. 25.1 b).[44]

The situation is more complex when the AI system is not an HRS and does not pursue the purposes of Annex III. In such cases, they would be systems outside the general framework of AIA obligations and would fall outside the scope of Article 6.4 AIA. However, Article 25.1(c) will clearly apply if the deployer of an AI system "changes the intended purpose" and "converts it into a high-risk AI system in accordance with Article 6."[45] In such a case, it will be considered as a provider and will have to comply with its obligations.

V. The role of the Commission, Criteria, Delegated Acts, updating and amendment of High-Risk schemes, in particular from Annex III.

Although the AIA has set criteria for the consideration of Annex III RAS, these are clearly subject to casuistry and interpretation. In recognition of this, significant power is given to the Commission to further delimit these cases both through guidelines and delegated acts. Furthermore, the AIA itself specifies to the Commission the parameters that should guide its action.

Thus, according to Art. 6.5 AIA, "the Commission [...] shall issue guidelines specifying the practical application of this Article in line with Article 96, together with an exhaustive list of practical examples of use cases of high-risk and non-high-risk AI systems." These guidelines should be placed within the scope of Article 96 AIA, whereby the Commission will take particular account of SMEs or local public authorities, as well as the state of the art, harmonized standards or technical specifications. The Commission will also "adopt delegated acts" to "amend" and even add "new conditions" to the HRS criteria linked to Annex III of Art. 6.3.1.

AI system is used in accordance with its intended purpose or reasonably foreseeable misuse, in particular where such risks persist despite the application of other requirements set out in this Section."

[44] "(b) when it substantially modifies a high-risk AI system that has already been placed on the market or put into service in such a way that it remains a high-risk AI system within the meaning of Article 6;".

[45] "(c) where it changes the intended purpose of an AI system, including a general-purpose AI system, which has not been considered as high risk and which has already been placed on the market or put into service, in such a way that the AI system in question becomes a high risk AI system in accordance with Article 6."

It is important to note that the starting point in these Commission actions is that if an AI system pursues Annex III purposes it is an HRS and only exceptionally will it not be an HRS (Whereas 52: "not exceptionally considered to be high risk"). Consequently, the Commission's action in determining criteria and delegated acts will have to be well justified, especially when it comes to not considering systems used for Annex III purposes as HRS. In this direction, the AIA contains some criteria to be followed by the Commission with regard to delegated acts (Art. 6.6):

- It may remove or modify the conditions for being considered as an HRS "only where there is concrete and reliable evidence of the existence of AI systems falling within the scope of Annex III, but which do not pose a significant risk of causing harm to health, safety or fundamental rights."

- Furthermore, "no amendment shall reduce the overall level of protection."

- It will take into account technological and market developments.

In addition, Article 7 regulates the "addition or modification" of Annex III use cases through these delegated acts. For this addition, the areas must be those already included in Annex III and a risk assessment must be carried out in which this article identifies many criteria to be taken into account by the Commission, such as: the specific purpose of the system, the extent of AI use, "the nature and amount of data processed, in particular if special categories of personal data are processed," "the degree of autonomy of the AI system and the possibility of a human being overriding a decision or recommendation," actual impacts on health, safety or fundamental rights or the likelihood of their occurrence according to documented reports or allegations, extent of this harm, "large number of individuals or disproportionate impact on a particular group of people," dependence these individuals may have on the outcome of that AI use, "imbalance of power and position of vulnerability," status, authority, knowledge, economic or social circumstances, or age of the individual concerned. Also, the possibility of "correcting or reversing the outcome" and the likelihood that the deployment of the AI system will be beneficial, as well as "the extent to which existing Union law provides for effective compensatory measures to prevent or significantly reduce those risks."

As noted above, it can be understood that these criteria to be followed by the Commission may also be criteria on which to rely in assessing whether a system is an Article 6 HRS.

In addition to the criteria for Commission delegated acts, the AIA also contains some operational mandates for the Commission to draw up annually "the list of high-risk AI systems." Thus, "it is of particular importance that the Commission carry out appropriate consultations during the preparatory phase, including at expert level" (Whereas173). The HRS list should be evaluated once a year (Whereas 173). This is relevant given that the general review of the Regulation should take place after five years and then every four years, and Annex III should be reviewed every four years (and two years after implementation of the AIA, Whereas 174).

VI. Summary and Conclusions

The EU has adopted a risk-based model for AI regulation, and the concept of HRS has become key. In my view, the risks that AIA defines are truly high risk, and their impact on people, rights and the democratic system is unquestionable. As discussed, the consideration of "high risk" is the central pillar of the AIA, and much of the regulation revolves around this concept. The essential elements of what is an HRS have been analyzed, underlining the importance that the regulation of an AI system as an HRS by the AIA does not imply its legal empowerment in terms of legal limits on fundamental rights, data protection, justice or other areas. This means that classification as a HRS does not exempt developers and users from having a specific legal basis for the use of such systems.

The AIA establishes a dual system for qualifying an HRS. On the one hand, AI systems that are safety products and components or Annex I products, specifically those that require a conformity assessment by an independent third party. These AI systems are HRS especially because of the hazards to integrity, safety and reliability. On the other hand, and arguably more importantly, Annex III of the AIA details up to twenty-five purposes that qualify an AI system as HRS. These include AI applications in biometrics, critical infrastructure, education, the workplace, essential services, emergencies, law enforcement, migration, asylum, border control, administration of justice and democratic processes. These purposes have been summarily described, and are sometimes discussed further in other sections of this book. The impact in this case is clearly linked to numerous fundamental rights.

It has been argued that there is no longer an automaticity whereby if the system has one of the 25 Annex III purposes it automatically becomes an HRS . The evolution of the AIA has incorporated the requirement that the AI system must have a substantial influence on decision-making. Thus, an AI system is considered to substantially influence decision-making if its results directly affect decisions that are important for Annex III purposes. Fortunately, this criterion has been regulated in a much more refined way than the solely automated decisions of Article 22 GDPR. This criterion is essential to determine the high-risk classification and thus whether the corresponding obligations apply.

It has been stressed that there is a presumption that any AI system pursuing Annex III purposes is an HRS, which entails special obligations and actions on the part of providers and users. This presumption warrants a level of caution, obligations on providers and additional controls, which could lead to problems in the future. It may be more *cost-effective* for a provider to do nothing, rather than to deny that its system is an HRS. It has also been specified that profiling with AI for the purposes of Annex III will always be HRS.

Finally, it was noted that the European Commission will play an essential role in updating and modifying the elements that define an HRS. This will be done through criteria and delegated acts, to ensure that the AIA remains up to date and effective in response to technological developments and market changes. Only time will tell whether the EU has gone too far in considering HRS and thus imposing a whole series of obligations.

THE REGULATION OF PREDICTIVE POLICING SYSTEMS IN THE ARTIFICIAL INTELLIGENCE ACT

Fernando Miró Llinares and Mario Santisteban Galarza

Fernando Miró Llinares. Professor of Criminal Law and Director of the CRIMINA Centre, Miguel Hernández University of Elche.[1]

Mario Santisteban Galarza. University of the Basque Country

I. Introduction

Before AI became a reality, some ideas, generally dystopian, about the use of these technologies in policing had already taken root in the collective imagination. Whether in the form of robotic police or systems that predict crime before it happens, such as those imagined by Philipp K. Dick in the story "The *Minority Report*", the culture had already made explicit some of the promises and risks that this technology could bring with it prior to technological development. It is understandable, therefore, that one of the first uses of AI systems to cause social concern was policing, especially when it became known that there were already algorithmic systems (some based on AI, others not) framed within the general concept of "predictive policing". The fact is that, mainly through data protection legislation, there was a tendency to limit certain police uses of algorithmic systems. High Courts have limited the massive processing of data on criminal offences because of the absence of appropriate safeguards[2], or because of the opacity of certain algorithmic systems[3]. The Court of Justice of the European Union has ruled that the systematic collection of biometric and genetic data in the framework of criminal

[1] The publication of this work is part of the Ius-Machina project on the normative bases and the real impact of the use of predictive algorithms in the judicial and penitentiary fields TED2021- 129356B-I00, funded by MCIN/AEI/10.13039/501100011033 and by the European Union "NextGenerationEU"/PRTREThis work is part of the GODAS* project: Project PID2022-137140OB-I00, funded by MCIN/AEI/10.13039/501100011033/FEDER, EU.

[2] This is the case of the German Constitutional Court, see in this sense Cotino Hueso, L., "Una regulación legal y de calidad para los análisis automatizados de datos o con inteligencia artificial. Los altos estándares que exigen el Tribunal constitucional alemán y otros tribunales, que no se cumplen ni de lejos en España", *Revista General de Derecho Administrativo*, *(*2023).

[3] It is well known that in the Netherlands the SyRI system, designed to detect fraud in certain aspects of the social security system, was declared contrary to Art. 8 of the European Convention on Human Rights, the lack of transparency of the system's parameters being of particular importance in the decision. See in this regard Appelman, N., Ó Fathaigh, R., and van Hoboken, J., "Social Welfare, Risk Profiling and Fundamental Rights: The Case of SyRI in the Netherlands", *JIPITEC* 257 12 (2021).

proceedings is unlawful[4]. The European Court of Human Rights[5] has limited the use of facial recognition technology by law enforcement authorities, making it necessary in a democratic society. Also, Directive 2016/680 of 27 April 2016 on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, provides safeguards against the processing of special categories of data in the framework of criminal law enforcement, and against automated decisions. However, its nature as a Directive and the breadth with which it approaches the issue of automation of decisions call for specific responses to the problems of AI in the field of criminal law enforcement[6].

Hence, in light of the European Union's decision to regulate AI, one of the areas subject to "preferential" regulation was the police use of these systems, particularly predictive ones. This paper addresses the regulation by the General Regulation on Artificial Intelligence of predictive AI systems used, or potentially usable, in the field of law enforcement and criminal justice. It does so in order to understand the regulatory provisions, to delimit their scope of application and to provide interpretative criteria for their application, but also to do so in relation to their current and potential police use. The intention is not only to understand which uses are declared prohibited and which are considered high-risk and what this implies, but also to frame all of this within the role that such systems, and others like them, currently play in policing. Thus, we will try to understand not only the regulatory implications from the perspective of which citizens' rights will not be affected due to the prohibition of some systems or the imposition of obligations for other systems considered high-risk, but also what this entails from the perspective of the reality of policing itself, which in recent years has been immersed in a strong trend towards technification and the reorientation of its functions towards prevention and prediction. Therefore, we will begin by framing the concept of predictive policing, proceeding to a simple categorisation of the

[4] "The mere fact that a person is being investigated for the commission of an intentional public offence cannot in itself be regarded as evidence from which it may be presumed that the collection of his biometric and genetic data is strictly necessary in view of the purposes for which it is intended and having regard to the infringements of fundamental rights, in particular the rights to respect for private life and to the protection of personal data guaranteed by Articles 7 and 8 of the Charter, which derive from it" CJEU of 26 January 2023 (Case C-205/21), para. 130.

[5] ECHR of 4 July 2023 (GLUKHIN v. RUSSIA, application no. 11519/20).

[6] Simón Castellano, P., "Inteligencia artificial y Administración de Justicia: ¿Quo vadis, justitia?", *IDP. Revista de Internet, Derecho y Política*, (2021), n.º 33.

algorithmic systems that fall within it and an explanation of which of them use AI and what singularities it provides, so that when we proceed to the analysis of the legal text we can frame it in relation to the reality that it is regulating and better understand its implications. The third section will be devoted to the evolution of the regulatory text in order to situate the final regulation in its context, and in the fourth section we will try to recapitulate, outlining some brief conclusions.

II. Predictive Policing Systems and Artificial Intelligence

1. Police organisation in times of digitalisation and the "misnamed" predictive policing

Policing has changed enormously over time. As a "permanent public organisation charged with the maintenance of security and order through the quasi-monopolistic exercise of state powers (basically coercion)"[7] the police, as we know them today, was born with the industrial revolution and cannot be explained without the emergence of the rule of law and the development of cities, but its functions and the way they are exercised have been changing due to multiple factors related to the political and social vision of what the exercise of control should be, or to the evolution of the type of problems it has been facing. Another driver of change in the institution and in police activity, has been, precisely, technology[8]. Perhaps because it has determined the possibilities of police action and, therefore, shaped the public's and the political powers' view of it, the fact is that technological changes have not only modified the way in which police action is carried out, but have also affected the functions of the police. The appearance of the radio and the automobile determined a more reactive police force than the previous one, less concerned with investigating crime and more focused on reacting to and controlling it; digitalisation and the process of "datification" in which we find ourselves directed the police function towards crime management and prevention[9], through the use of crime-related information and the management of resources to use social control to try to prevent crime from occurring. Within this stage, the emergence of Big Data and AI systems seems to lead

[7] Guillén Lasierra, F., *Modelos de policía. Towards a plural security model,* Barcelona, Bosch, (2016).

[8] Deflem, M., and Chicoine, S. "History of technology in policing. *J Psychopharmacol*, 24 (2) (1988), pp. 141-145.

[9] González-Álvarez, J., Santos Hermoso, J., and Camacho-Collados, M., "Predictive policing in Spain. Application and future challenges", *Behavior & Law journal*, 6(1), (2020).

police action towards the automation of preventive work under the idea of "prediction".

Indeed, the type of police practice that comes under the heading of "predictive policing" would only be understood in the context of the digitalisation, "scientisation", bureaucratisation and automation of police work that has taken place in recent decades. In fact, the first use of the term is attributed to William Bratton, the father of Compstat, the crime statistics system which, building on the foundations of Goldstein's Problem Oriented Policing and the computerisation of police work begun in the 1960s and 1970s, developed in the late 1980s a statistical system for computerised integration of police data[10] which, over time, as Wilson has pointed out, accelerated data collection and processing, determined the integration of computerisation into routine patrol work on an unprecedented scale, significantly increased the use of crime mapping techniques to analyse the geographical distribution of patrols and to locate "times of day" and "crime peaks" in known "hot spots" and, supported by the academic development of so-called "environmental criminology", later led to the development of crime analysis techniques such as Intelligence Led Policing or Evidence Based Policing that advocate data-driven decision making and strategic problem solving for police management, resource allocation and crime control. But is this predictive policing or is it something more, and how does it all relate to AI?

The term predictive policing began to become widespread in academia in the early 2010s to refer to a set of "analytical techniques -particularly quantitative techniques- that, by making statistical predictions, seek either to identify potential targets for police intervention and prevent crime or to solve past crimes"[11]. Similar is the definition proposed by Ratcliffe who, in addition to preferring the term "crime forecasting"[12], defines it as "the use of historical data to create forecasts of crime areas or crime hotspots, or characteristic profiles of high-risk offenders that will be a component of police resource allocation decisions"[13]. In recent years the term predictive policing has begun to be abandoned by its main users, US police departments[14]. But a new brand

[10] Miró Llinares, F., "Predictive Policing: Utopia or Dystopia? On attitudes towards the use of Big Data algorithms for law enforcement", *IDP. Internet, Law and Policy Journal*, n.º 30., (2020).

[11] Walter, P., McInnis, B., Price, C., Smith, S., and and Hollywood, J., *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*, Santa Monica, CA: RAND Corporation, (2013).

[12] Ratcliffe, J. "Predictive policing", in Weisburd, D and Braga, A.A. (EDS.), Police *innovation. Contrasting perspectives*. 2nd. edition. Cambridge: Cambridge University Press (2019).

[13] Ibid.

[14] Especially since the Los Angeles Police Department in the spring of 2020 (followed

name, "data driven policing", is being used, after which the use of some of these tools remains in effect. Perhaps because the promises of increased efficiency in decision-making, reduction of bias and subjectivity in decision-making, etc., and of technological solutionism are not only still valid but are being augmented by the generalisation of AI systems[15].

While Wilson is right to point out that within the broad term predictive policing we find a "medley of contemporary policing methodologies and philosophies" that link digitisation and the trend towards automation in decision-making with an idea of prediction[16], there is agreement in highlighting two types of crime forecasting techniques, crime forecasting focused on where the crime is going to be committed, and crime forecasting focused on the people who are going to commit or be victims of the crime. Both techniques can be carried out with or without the use of AI systems, something we will return to later when we discuss the ethical risks associated with them. Both techniques are also characterised by the use of historical crime data to identify targets of police interest for the purpose of preventing crime, reducing the risk of crime or disrupting criminal activity[17], but they differ in the type of input they are fed and the output they generate. Place Based Predictive Policing techniques focus on determining where and when crimes of each type are perpetrated and extract patterns of risk in such environments to improve decision-making on where and when to intervene preventively. Person Based Predictive Policing techniques[18] of perpetrators (offenders)[19] or victims (victims), establish general patterns or profiles of offenders and estimate who and when they are most likely to re-offend or become victims

by many others thereafter) stopped using the "predpol" system because of criticisms of racial discrimination against these tools, which increased with the BLM movement. Davis, J., Purves, D., Gilbert, J., & Sturm, S. (2022). Five ethical challenges facing data-driven policing. *AI and Ethics*, *2*(1), 185-198.

[15] López Riba, JM., "Inteligencia artificial y control policial. Cuestiones para un debate criminológico frente al hype", in press, (2024).

[16] Wilson, D. "Predictive policing management: A brief history of patrol automation", *New formations*, *98*(98), (2019). 139-155.

[17] Ratcliffe, J. "Predictive policing", in Weisburd, D and Braga, A.A. (EDS.), Police *innovation. Contrasting perspectives*. 2nd. edition. Cambridge: Cambridge University Press (2019).

[18] They have also been referred to as Ofender Focused Crime Forecasting tools, although in this case they refer exclusively to the estimation of the likelihood of crime perpetration and do not include victimisation.

[19] On heat lists and similar systems for calculating the risk of individual persons such as Beware, see Degeling, M. and Berendt, B., "What is wrong about Robocops as consultants? A technology-centric critique of predictive policing AI and Society", 33 (3) (2018). pp. 347-356.

of crime[20]. These techniques have common elements: a) they seek to estimate the likelihood of a crime being perpetrated, b) they use data from crimes already committed and, therefore, information from criminals and victims, c) they are computer-based decision support tools, and d) they are predefined algorithms and may or may not use machine learning algorithms, although most do not. But they also have very relevant distinguishing features: 1) the scientific bases of each are very different (environmental criminology in the case of PlaceBPPs[21] and, generally, violence risk assessment in the case of PersonBPPs[22]; 2) place-based tools only take into account data on the offender or victim that relate to the type of offence, and the place where it has been perpetrated, but do not profile potential victims or offenders based on the generalisation of those who perpetrate (or suffer from) such crimes; 3) they lead to very different types of policing.

Alongside these two techniques that can be considered as police crime prediction techniques or, wrongly called, predictive policing, there is a third and final set of police techniques that are sometimes considered to be included in the same category and are all those techniques based on surveillance by means of images and which, based on facial recognition techniques, movement recognition, number plate reading, etc., are combined with algorithms to identify suspicious subjects and "predict" possible criminal actions[23]. Police organisations are increasingly using facial recognition software that it is fed by traffic cameras, body cameras worn by police officers themselves, or number plate cameras and similar devices, that produce digital data that can be combined to identify and track individuals for security and protection[24]. Although it is said that these techniques could be used to predict the occurrence of unlawful conduct[25], the fact is that they are techniques for investigating crimes already committed or in the process of being committed, rather than for

[20] This is the case of Viogen, Presno Linera, M. A., "Policía predictiva y prevención de la violencia de género: el sistema VioGén" IDP. Revista de Internet, Derecho y Política, 2023, n.º 39, pp. 1-13.

[21] See, for example, FELSON, M. "Routine activities and crime prevention in the developing metropolis", *Criminology*, vol. 25, (1987).

[22] González-Álvarez, J., Santos Hermoso, J., and Camacho-Collados, M., "Predictive policing in Spain. Application and future challenges", *Behavior & Law journal*, 6(1), (2020).

[23] Miró Llinares, F., "Predictive Policing: Utopia or Dystopia? On attitudes towards the use of Big Data algorithms for law enforcement", *IDP. Internet, Law and Policy Journal*, n.º 30., (2020).

[24] Hannah-Moffat, K. "Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates", *Theoretical Criminology*, *23*(4), (2019), 453-470.

[25] Ferguson, A. G. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. NY: NYU Press, (2017).

estimating the probability of crimes being committed. Their consideration as predictive policing stems from the consideration that this wide-ranging toolbox also includes the use of techniques to "predict" the identity of the perpetrators of crimes[26]. In any case, the analysis of these tools, which fall more into the logic of crime investigation than crime prevention/prediction, will be left out of this chapter as they are closely related to facial recognition techniques, which have received special treatment by the AIA as the ethical risks they pose are different.

2. The ethical risks of predictive policing (and those added by the use of Artificial Intelligence)

In this process of bureaucratisation and digitalisation of the police, AI is presented as the latest exponent of improving effectiveness and efficiency (and even reducing subjectivity) in the exercise of its tasks, in general, and in crime prevention in particular. Big Data and the emergence of new computational techniques, such as machine learning, open up the possibility of improving estimates that were previously made without so much data or without these techniques. The use of AI would improve the accuracy of estimates, optimise the distribution of resources, and therefore reduce the costs of police action[27], even ensuring more equitable action[28]. In fact, these ideas seem to be in line with what has been established by the AIA itself, which in recital 4 notes that "by improving prediction, optimising operations and resource allocation, and personalising digital solutions available for individuals and organisations, the use of AI can provide key competitive advantages to undertakings and support socially and environmentally beneficial outcomes" from many points of view, among which it expressly cites security. But the same regulation, in recital 5, recognises that AI systems can "generate risks and cause harm to public interests and fundamental rights that are protected by Union law". It is therefore necessary to identify what these risks may be

[26] Brayne, S., Rosenblat, A., and Boyd, D., *Predictive policing. Data & civil rights: a new era of policing and justice*, (2015). https://www.datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf

[27] Brayne, S., Rosenblat, A., and Boyd, D., *Predictive policing. Data & civil rights: a new era of policing and justice*, (2015). https://www.datacivilrights.org/pubs/2015-1027/Predictive_Policing.pdf

[28] Saphiro calls this narrative a technopolitical tactic used by police agencies "predictive policing for reform", a misguided attempt to rationalise police patrols by algorithmically restructuring their actions. Saphiro, A., "Predictive Policing for Reform? Indeterminacy and Intervention in Big Data Policing", *Surveillance & Society*, 17(3/4), 2019, pp. 456-472.

in order to be able to assess the approach of the European regulation, also in the area of the use of predictive policing systems.

In our opinion, there are two levels that must be taken into account in a differentiated way, firstly, and then connected, when assigning possible ethical risks for the police use of so-called predictive systems: on the one hand, there is the level of the prognostic technique (either of place or of person) that is behind the tool, and on the other, the algorithmic technique (either classical or automated or AI) that is used at the computational level. Each of these dimensions or levels makes it possible to identify different techniques, but also risks to their own distinctive rights and guarantees, which, moreover, can be combined in different ways.

Firstly, depending on the prognostic technique used by the tool (either location or person), although there will be shared ethical problems or risks, others will be different or specific to only some of them due to a) the different nature of the data used, b) their different scientific logic, or c) the different nature of the impact of predictive policing in one or the other case. On the question of data, its quality, and the potential biases it can produce, police systems that seek to predict where crimes will be perpetrated are often informed by data from complaints, although they can also be informed by data from crimes investigated by the police. This type of data can be affected by different biases related to the greater police presence in those areas, the way in which data is collected by the police, the crime events that are most reported, among many others[29]. In these cases, and we will see that this is important for AIA, we cannot say that patterns of people are being carried out on the basis of the generic characteristics associated with them, since it is the places that are the object of analysis. But that does not mean that possible biases do not end up affecting people indirectly. Thus, the fact that more patrols are carried out in certain areas can lead to a greater number of interventions and police stops and, therefore, determine a greater number of arrests and criminal offences that end up affecting the people who live in those areas as they are more exposed to police surveillance by the algorithms built on such information.

In the case of police systems that seek to predict the likelihood of a person committing or being victimised by crime or to improve related decision-making, the data on which such algorithms are based do include multiple variables related to personal characteristics associated to age, criminal history, gender, and other similar factors that relate to crime perpetration or victimisation.

[29] See on all this, with multiple references and details, López Riba, JM., "Inteligencia artificial y control policial. Cuestiones para un debate criminológico frente al hype", in press, 2024.

In addition to the issue of data, whether the prediction relates to the crime scene or to the people involved in the crime raises other ethical issues. One of them has to do with the validity and reliability of the predictions, not because the predictions in one case are generally better than those in the other, but because the scientific bases for the two systems are different. PlaceBPP tools are based on the premises of crime geography and geographical crime analysis techniques, and have shown high reliability in some experiments[30], although it is more contested whether they are able to prevent and reduce crime beyond simply predicting police intervention[31]. There are also other ethical problems related to the application of these tools, such as the fact that they change the way in which police officers act and stop performing essential community functions to become detectives of places of risk[32], or that they confuse prognosis with intervention[33]. In the case of PersonBPPs, the assessment of risk of violence as an alternative method to the diagnosis of dangerousness for the prediction of violence[34], which is based on knowledge of the risk factors associated with violence, identifying the causes that explain and the factors that are related to violent behaviour (risk factors) and also those that influence the reduction or abandonment of violent and/or criminal activity (protective factors), which may be common or specific to different forms of violence which, in turn, may be more or less related to criminal *behaviour*[35]. The use of these actuarial instruments and actuarially based structured clinical judgements has increased significantly in the police field, but also in the judicial field for decision making related to the estimation of the risk of recidivism or of breaking a sentence, based on the consideration of their high predictive capacity[36] which, however, can and should be qualified. Martínez Garay has pointed out that these tools achieve probably satisfactory levels of

[30] Ratcliffe, J. "Predictive policing", in Weisburd, D and Braga, A.A. (EDS.), *Police innovation. Contrasting perspectives.* 2nd. edition. Cambridge: Cambridge University Press (2019).

[31] Miró Llinares, F., "Predictive Policing: Utopia or Dystopia? On attitudes towards the use of Big Data algorithms for law enforcement", *IDP. Internet, Law and Policy Journal*, n.º 30., (2020).

Also, López Riba, JM., "Inteligencia artificial y control policial. Cuestiones para un debate criminológico frente al hype", in press, (2024).

[32] Ferguson, A. G. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement.* NY: NYU Press, (2017).

[33] Ibid.

[34] Andres Pueyo, A. and Illescas, S. R.. "Predicting violence: between dangerousness and the assessment of the risk of violence". *Papeles del psicólogo*, *28*(3), (2007), 157-173.

[35] Ibid.

[36] Andres Pueyo, A. & Illescas, S. R. "Predicting violence: between dangerousness and violence risk assessment". *Papeles del psicólogo*, *28*(3), (2007), 157-173, also, Skeem, J. L., & Mo-

precision in estimating the relative risk of recidivism, but are more imprecise in estimating the absolute risk of recidivism, and produce an overestimation of risk when applied to phenomena with low prevalence rates such as violent crime[37]. This being so, what is relevant is to understand that it is not a matter of dichotomising whether these tools predict or do not predict, but rather of understanding which of the existing tools available provide better comparative criteria for judicial decision-making on dangerousness[38], and also which are the real limitations of each system to carry out each specific assessment for each type of behaviour[39].

And this must be linked to the other fundamental element that must be taken into consideration when assessing the ethical risks associated with either type of prediction: the normative consequences of policing in either case. While place-based predictive policing techniques essentially serve to make consistent decisions on the allocation of police resources and, therefore, its consequences will consist of there being geographical areas more or less monitored and with more or less police presence, personal forecasting techniques can have consequences directly related to fundamental rights. As Martínez Garay has pointed out, generating expectations about the real possibilities of predicting violent behaviour is more dangerous where the consequences are the possible infringement of citizens' freedom[40]. This can also lead us to differentiate between the ethical problems associated with tools that estimate the commission of crimes versus those that estimate victimisation. Although it may be reasonable to think that the latter pose fewer problems, the truth is that both are, in reality, generally based on the assessment of the risk of violence based on the conduct of a potential aggressor, so it will be necessary to consider whether the measures of one and the other estimates

nahan, J. "Current directions in violence risk assessment. Current Directions in Psychological Science", 20(1), (2011), pp. 38-42.

[37] Martínez Garay, L. "Conceptual errors in the estimation of recidivism risk: The importance of differentiating sensitivity and predictive value, and absolute and relative risk estimates". *Spanish Journal of Criminological Research: REIC*, (14), (2016).

[38] Miró Llinares, F. and Castro Toledo, F. J., "Correlation does not imply causality? El valor de las predicciones algorítmicas en el sistema penal a propósito del debate epistemológico sobre el fin de la teoría"' in Demetrio, E., de la Cuerda Martín*,* M. and García de la Torre García, F., *Derecho penal y comportamiento humano. Avances desde la neurociencia y la inteligencia artificial*, Tirant lo Blanch, 2022.

[39] Martínez Garay, L., & Montes Suay, F., "El uso de valoraciones del riesgo de violencia en Derecho Penal: algunas cautelas necesarias", *InDret: Revista para el análisis del derecho, 2018*, N. *2/2018,* (2018)*,* 1-46.

[40] Ibid.

are different or whether, by the fact of being applied to the victim for their protection, the measures will also have an impact on the potential aggressor.

Beyond the type of prognosis carried out by each tool or technique, there is a second level that must be taken into account when assessing the ethical risks posed by systems that, too generically, are often grouped under predictive policing. This concerns the computational technique used and, in particular, whether or not the system is AI-based, meaning, as stated in Article 3(a) of the Regulation, "a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments"[41]. The truth is that predictive policing systems were born before the popularisation of AI and most of them were developed without the use of automated learning techniques such as machine learning. It was only with the spread of Big Data and the possibility of accessing large amounts of information that the possibility of using these techniques to improve estimates for both location and person prediction systems began to be considered. But why, in terms of ethical risks, is it relevant whether or not predictive systems use machine learning or similar computational techniques that can indicate we are dealing with AI due to a certain degree of independence from human intervention? Or, put another way, what does the use of AI add in terms of ethical risk? In our view, there are three fundamental questions: the different scientific logic behind AI algorithms; the question of the traceability and explainability of decisions; and, related to all this and in the background, the problem of the autonomy of AI.

Starting with the different scientific logic of the two systems, the classic predictive algorithms that emerged in the midst of digitisation used large official datasets. But these had been deliberately configured by social science researchers, following methodological guidelines and from specific theoretical frameworks based on a causal scientific logic to estimate the risk of vio-

[41] As stated in recital 12 of the Regulation, this implies not including within these systems other simpler traditional software, and not including systems that rely solely on rules defined only by natural persons to execute operations automatically. The recital also states that "Techniques that enable inference while building an AI system include machine learning approaches that learn from data how to achieve certain objectives, and logic- and knowledge-based approaches that infer from encoded knowledge or symbolic representation of the task to be solved", and also that the fact that AI systems are designed to operate with different levels of autonomy 'meaning that they have some degree of independence of actions from human involvement and of capabilities to operate without human intervention'. We will return to this later.

lence or crime patterns, depending on the type of tool[42]. The advent of big data analytics was a radical change, not only because of the large amounts of information that can be accessed but also because it allows, and even demands, the use of new techniques that are designed from a different logic in which causal inference is not relevant and the correlation between factors is everything[43]. Unlike traditional predictive algorithms, these new algorithms start from large amounts of big data; they are able to work with real-time data and to adapt thanks to machine learning; they are constrained neither in the data they collect nor in the results they produce by pre-determined theoretical frameworks and, indeed, they need not be implicitly designed to predict where crime will occur or whether someone will commit the crime, but are capable of contrasting multiple variables about individuals, places and societies to make supposedly more accurate predictions about very different elements[44]. The results of these algorithms do not attempt to explain why someone will commit a crime or where it will occur, but to estimate it regardless of the variables leading to the forecast and the causal meaning of these variables. These algorithms may therefore have fewer problems of producing biases related to the selection of variables by social researchers who have a particular view of the problem, but they may carry the risk of not being able to explain the meaning of the estimates.

In fact, closely related to this, it has been said that AI brings with it more problems of decision traceability for two reasons. The first is the issue of the opacity of algorithms, although the so-called "black box" problem is not unique to AI and can also be associated with traditional actuarial or location prediction tools. Many of these algorithms are not accessible to the public and have intellectual property rights that prevent access to them, so that the reasons for the decisions could not be followed, raising problems of lack of transparency and, in particular, the enormous risk that the right to defence cannot be properly exercised. Secondly, AI algorithms, even if traceable, are fluid and transformative, changing unpredictably based on the new data they

[42] Hannah-Moffat, K., 'Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates', *Theoretical Criminology*, *23*(4), (2019), 453-470.

[43] Miró Llinares, F. and Castro Toledo, F. J., "Correlation does not imply causation? El valor de las predicciones algorítmicas en el sistema penal a propósito del debate epistemológico sobre "el fin de la teoría"" in Demetrop, E., de la Cuerda Martín*,* M. and García de la Torre García, F., *Derecho penal y comportamiento humano. Avances desde la neurociencia y la inteligencia artificial*, Tirant lo Blanch, (2022).

[44] Hannah-Moffat, K., 'Algorithmic risk governance: Big data analytics, race and information activism in criminal justice debates', *Theoretical Criminology*, *23*(4), (2019), 453-470.

introduce[45] and, in the case of tools using deep learning, giving rise to solutions that are difficult to explain in causal terms.

And this, in turn, is related to the last of the characteristics that adds, in terms of risk, uniqueness, to predictive algorithms that use AI: the possibility of algorithms acting autonomously, even if it is in the learning of data. In addition to the functioning of some automated learning systems, such as deep learning, in which the absence of training and the difficulty of determining the variables involved already make it difficult to explain (understand) the origin of the decisions, some of them can "autonomously" decide how to select certain variables without being supervised and, therefore, can be avoided by a human[46]. Obviously, the risks involved are very different. Traditional prognostic systems are intended as guides or support tools for professionals, which do not replace them in decision-making, but inform them, and those who apply them are supposed to be trained to understand the logic of the recommendations. In AI systems, the fact that the recommendation is not fully explainable may lead to an automation of its application by the subject, which entails particularly relevant risks.

[45] Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., De Paor, A., Felzmann, H., Haklay, M., Khoo, S.-M., Morison, J., Murphy, M. H., O'Brolchain, N., Schafer, B., & Shankar, K. "Algorithmic governance: Developing a research agenda through the power of collective intelligence", *Big Data & Society*, 4(2), (2017).

[46] Closely related to the above, the CJEU has been reluctant to accept the use of machine learning in certain police uses. This is the case of the use of airline passenger name record data to detect terrorist offences and serious crimes provided for by Directive 2016/681, which establishes, in the words of the CJEU, "a continuous, non-selective and systematic surveillance regime, including the automated assessment of personal data of all persons using air transport services". The comparison of PNR data with the relevant databases is carried out on the basis of specific criteria, which precludes "the use of Artificial Intelligence technologies in the context of machine learning systems, which may alter, without human intervention and control, the evaluation process and, in particular, the evaluation criteria on which the result of the application of the process is based, as well as the weighting of those criteria". In this regard, the CJEU clarifies that "the use of these technologies would risk depriving the individualised review of positive results and the control of lawfulness required by the provisions of the PNR Directive of any useful effect. Indeed, as the Advocate General states, in essence, in point 228 of his Opinion, in view of the opacity which characterises the operation of Artificial Intelligence technologies, it may be impossible to understand the reason why a given programme has achieved a positive match. In these circumstances, the use of such technologies could also deprive the persons concerned of their right to effective judicial protection, which is enshrined in Article 47 of the Charter and which the PNR Directive seeks, according to recital 28 thereof, to guarantee at a high level, in particular in order to challenge the non-discriminatory nature of the results obtained".

III. Regulating predictive policing in the Artificial Intelligence Act

1. Developments in the regulation of the use of predictive policing systems in the legislative process of the Artificial Intelligence Act

The regulation of predictive policing has undergone substantial changes in the legislative procedure, demonstrating on the one hand the complexity of the task facing the community legislator and, on the other, the particularly controversial nature of this matter.

Initially, the European Commission's Proposal for a Regulation of the European Parliament and of the Council, presented in April 2021, did not outlaw the use of predictive policing tools. Among the prohibited practices mentioned in Title II, except for real-time remote biometric identification, which is not the subject of this chapter, none of them covered the use of AI systems to assess the risk of committing crimes. On the other hand, and by reference to Art. 6.2 of the text, the use of some of the tools described above could fit into the practices of Annex III, and thus constitute a high-risk use of AI. This Annex included certain applications "related to law enforcement". Among these, paragraph 6(a) mentioned "AI systems intended to be used by or on behalf of law enforcement authorities, or by Union institutions, bodies, offices or agencies in support of law enforcement authorities or on their behalf to assess the risk of a natural person becoming the victim of criminal offences". This would cover certain predictive policing systems already described, irrespective of the stage of the criminal law enforcement procedure and informing both police and judicial decisions.

In addition, Annex III also included as high-risk AI those "AI systems intended to be used by or on behalf of law enforcement authorities or by Union institutions, bodies, offices or agencies in support of law enforcement authorities for the profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680 in the course of the detection, investigation or prosecution of criminal offences" (6. e)). We will come back to this definition as it seems to be key in determining what is prohibited in the final text. For the moment, what we are interested in highlighting is the similarity between the two practices described (letters a) and e)), it being very difficult to differentiate "the prediction of the frequency or repetition of a criminal offence" (letter 6 e)), from the assessment of the risk of the commission of the offence or criminal recidivism (letter a)). Closely linked to this, the carrying out of "individual risk assessments of natural persons with the aim of determining the risk of them committing criminal offences", entails, in practice, the drawing up of profiles of suspected criminals in the terms set out in the aforementioned Directive (EU) 2016/680, which again blurs the differentiation of the factual assumptions referred to in the rule in both sections. Surely the difficulty of differentiating between the two cases of fact has led subsequent versions to dispense with this division.

Finally, in paragraph 6(g), a high-risk use was introduced with, in this case, a clearly differentiated factual scenario: 'AI systems intended to be used to carry out criminal offence analysis in relation to natural persons to enable law enforcement authorities to examine large sets of complex linked and unlinked data available in different sources or formats to detect unknown patterns or discover hidden relationships in the data'. In this case, the Commission's proposal seemed to cover PlaceBPP techniques as the assessment process, does not aim at assessing the risk posed by a subject, irrespective of the fact that this assessment is fed by data relating to criminal offences that do contain personal data[47].

Some relevant changes were already made with the Council's Common Position[48]. In particular, it introduced some relevant changes to the definition of these practices and also to the list of high-risk systems in Annex III. In what we are interested in highlighting here, point g) of paragraph 6 was eliminated, that is, what the Council's Common Position identified as systems for crime analytics, and which we have already pointed out can include location-based predictive policing systems. On the other hand, points a) and e) were maintained, although it was clarified that these tools were considered high-risk when used by law enforcement authorities, but also by other subjects delegated by these authorities.

The substantial change in the legal status of these tools is to be found in the European Parliament's amendments. Indeed, the legislator had already been wary of the use of these tools, stating in a resolution of 2021, "that while predictive policing can analyse the data sets necessary for the determination of patterns and correlations, it cannot answer the question of causality and cannot make reliable predictions of individual behaviour, and therefore

[47] On the other hand, mention should be made of other systems which, while not strictly speaking "predictive policing", do fall within the scope of "predictive" or "evidence based" sentencing, and which were considered a high-risk system by the Commission's text. Article 8.a referred in particular to those AI systems intended "to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts" (8. a)). Consequently, all forms of *evidence based sentencing* fall, in any case, and to the extent that they do not fall into the above categories, into the group of high-risk systems.

[48] Available at: https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf

cannot form the sole basis for intervention"[49]. However, in the aforementioned decision, "the use of AI by law enforcement authorities to make behavioural predictions concerning individuals or groups on the basis of historical data and past behaviour, group membership, location or any other such characteristics, in an attempt to identify persons likely to commit a crime" was already opposed.

This distrust of predictive policing was strongly reflected in the amendments tabled[50]. The most relevant change with respect to the Commission's text consisted in the inclusion in the prohibited uses of Art. 5 of some techniques that would fit within the broad box of predictive policing, moving from high-risk use to prohibition. Thus, Parliament's amendment 224 elevated the following AI practice to the category of prohibited use: the placing on the market, the putting into service for this specific purpose, or the use of an AI system for making risk assessments of natural persons in order to assess or predict the risk of a natural person committing a criminal offence, based solely on the profiling of a natural person or on assessing their personality traits and characteristics, in particular the location of the person or past criminal conduct of natural persons or groups of natural persons (new Art. 5. 1 d).

The prohibited practice outlawed risk assessment systems, irrespective of the stage of processing at which they take place, and which involve profiling of the individual. It seems, in fact, that the Parliament's amendments closely followed data protection law, which sets relevant limitations to processing operations that constitute profiling. This is logical in view of the "potential discriminatory effects on natural persons on grounds of race or ethnic origin, political opinions, religion or belief, trade union membership, genetic

[49] European Parliament resolution of 6 October 2021 on Artificial Intelligence in criminal law and its use by law enforcement authorities in criminal matters (2020/2016(INI)).

[50] The position of the ECDC-SEPD in its Joint Opinion 5/2021 was no doubt also a contributing factor

on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) of 18 June 2021. On policing systems they stated: "The determination or classification by a computer of future behaviour independently of one's own will also affects human dignity. AI systems intended to be used by law enforcement authorities to carry out individual risk assessments of natural persons for the purpose of determining the risk of their committing criminal offences (see Annex III(6)(a)), or for predicting the frequency or repetition of an actual or potential criminal offence on the basis of profiling of natural persons or assessment of personality traits and characteristics or past criminal behaviour (see Annex III(6)(e)) used for their intended purpose will lead to the fundamental domination of law enforcement and judicial decision-making, with the consequent reification of the individual concerned. Such AI systems, which go to the core of the right to human dignity, should be prohibited under Article 5.

condition or health status or sexual orientation"[51] that automated decisions based on profiling may cause[52]. Thus, the recitals of the text amended by the Parliament argue that these systems "entail a particular risk of discrimination against certain persons or groups of persons, as they violate human dignity as well as the key legal principle of the presumption of innocence" (recital 26a, amendment 50).

On the other hand, as far as high-risk systems are concerned, contrary to the Council's common position, the inclusion of point (g) is maintained, and PlaceBPP systems are to be considered as remaining in this category.

The final text closely follows the European Parliament's dual approach, qualifying some uses of predictive policing tools as prohibited practices and others as high-risk uses. However, probably because of the divergence of views between the Council and the Parliament, manifested in the significant differences between the Council's common position and the Parliament's amendments, the final text has also softened its response to this use of AI.

The clearest example of this "trade off" between the Parliament and the Council is the description of the prohibited practice concerning PersonBPP. We will analyse it below by distinguishing between the two main measures that the Regulation finally adopts in relation to predictive systems: a) the prohibition of some of them; b) the consideration of other police crime prediction techniques as high-risk systems.

2. Predictive policing systems prohibited in the AI Act

Article 5.1 (d) of the AIA prohibits "the placing on the market, the putting into service for this specific purpose, or the use of an AI system for making risk assessments of natural persons in order to assess or predict the risk of a natural person committing a criminal offence, based solely on the profiling of a natural person or on assessing their personality traits and characteristics". The prohibition broadly respects the Parliament's text, but with the important qualification that the risk assessment is based solely on the profiling process.

It is also interesting to note that the object of the assessment has been limited, referring exclusively to a criminal offence, excluding administrative offences. On the contrary, the current wording allows the assessment to take

[51] CJEU of 7 December 2023, Case C-634/21, (OQ and Land Hessen).

[52] Also Recommendation CM/Rec (2010)13 of the Committee of Ministers to Member States on the protection of individuals with regard to automatic processing of personal data in the context of profiling noted that "such profiling may expose individuals to particularly high risks of discrimination and violations of their personal rights and dignity".

place at different stages of the proceedings and by different bodies, which seems to admit cases of predictive justice or sentencing. The reference to "the location of the person or the past criminal conduct of natural persons or groups of natural persons" as examples of variables to be taken into account in prediction is also deleted. This is not problematic as these characteristics can be taken into account in profiling, which, when present in the predictive process, determines the application of the prohibition; it should be noted that profiling and the assessment of personality traits and characteristics are alternative and not cumulative requirements.

As mentioned above, the prohibition covers assessments that are based "solely" on profiling. The AIA refers to the concepts of the general data protection law to define the prohibition. The GDPR defines profiling as "any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements" (Art. 4.4 of the GDPR). As indicated by the Article 29 Working Party, three requirements must be met for profiling to take place: it must be an automated form of processing; it must be carried out in respect of personal data; and the purpose of profiling must be to evaluate personal aspects relating to a natural person[53]. In view of the above, the breadth of the prohibition can be appreciated. It can hardly be argued that predictive systems based on the characteristics of the subjects (PersonBPP) do not involve profiling: they involve automated processing, they relate to information on an identifiable natural person, and their purpose is to assess personal aspects of the data subject (in this case his or her dangerousness)[54].

A possible justification for the Regulation to cover assessments based solely on profiling can be found in the recitals: "in line with the presumption of innocence, natural persons in the Union should always be judged on their actual behaviour". However, "persons should never be judged on AI-predicted behaviour based solely on their profiling, personality traits or characteristics, such as nationality, place of birth, place of residence, number of children, level of debt or type of car, without a reasonable suspicion of that person being involved in a criminal activity based on objective verifiable

[53] Article 29 Data Protection Working Party, *Guidelines on automated individual decisions and profiling for the purposes of Regulation 2016/679*, 6 February 2018.

[54] O., Lynskey, "Criminal justice profiling and EU data protection law: precarious protection from predictive policing", *International Journal of Law in Context*. 15(2), (2019), pp. 162-176.

facts and without human assessment thereof" (Recital 42). Indeed, profiling carries the risk of "generic correlations which may not be correct for all persons", treating a person as a member of a group rather than as an individual[55]. In fact, a generalisation is inherent to profiling in that its mission is to assign a person to a profile in the configuration of which not all relevant features of the individual will ever be represented[56]. Profiling, on the other hand, has social benefits, in the sense that it allows us to automate and lighten our understanding of the world, and there are parallels between algorithmic automation and biological automation[57]. However, in particularly sensitive areas, where individualisation of decisions is required, it is logical for the legislator to veto its implementation. In fact, in other cases, also in the area of criminal law enforcement, the legislator has prohibited that the sole criterion for making binding decisions is the ascription to a profile[58]. The AIA joins them in the understanding that the assessment of the risk of committing a criminal offence carried out exclusively in an automated manner and based on profiling entails unacceptable risks.

Having clarified the *ratio legis*, the second question to be analysed is when is the time in which the assessment should be understood to be based exclusively on profiling, which is a prerequisite for the prohibition. Once again, it seems essential to turn to data protection law, as both the GDPR and Directive 2016/680 use a very similar concept when proscribing decisions "based solely" on automated processing of personal data (Articles 22.1[59] and 11.1 respectively). Thus, the GDPR postulates that 'every data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her'. The scope of art. 22 is certainly controversial, and it would be foolhardy to postulate a conclusive interpretation of it. On this point, we simply follow the guidelines of the Article 29

[55] FRA, *Guidance on preventing unlawful profiling now and in the future*, 2019, Available at: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-preventing-unlawful-profiling-guide_es.pdf

[56] Palma Ortigosa, A., *Decisiones automatizadas y protección de datos. Especial atención a los sistemas de inteligencia artificial*, Dykinson, Madrid, (2022).

[57] Hildebrandt, M., "Defining Profiling: A New Type of Knowledge?", in Hildebrandt, M., and Gutwirth S., *Profiling the European citizen*, Springer, (2008).

[58] This is the case of Royal Decree 190/1996 of 9 February 1996, approving the Prison Regulations, which in Article 6.1 states that "No decision by the prison administration involving an assessment of the human behaviour of inmates may be based exclusively on the automated processing of data or information that provides a definition of the profile or personality of the inmate".

[59] In detail, on Art. 22, the recent CJEU of 7 December 2023 (Case C-634/21).

Working Party, which in interpreting art. 22 has pointed out that the fact that a decision is based solely on automated processing "means that there is no human involvement in the decision-making process". To this it adds that "to be considered as human involvement, the controller must ensure that any monitoring of the decision is meaningful, rather than merely a token gesture" and "must be carried out by a person authorised and competent to modify the decision"[60].

Therefore, among the different ways of introducing human supervision in the algorithm's lifecycle, the GDPR in its art. 22.1 establishes, at the heart of the prohibition, that a subject with real capacity to decide can modify the automated decision (human in the loop)[61]. In a very similar way, an exception to the prohibition of predictive policing is attached to art. 5.1 (d) of the AIA, which refers to the fact that "this prohibition shall not apply to AI systems used to support the human assessment of the involvement of a person in a criminal activity, which is already based on objective and verifiable facts directly linked to a criminal activity". In reality, this second subparagraph can be interpreted as a clarification rather than an exception. It tells us when a decision is not taken exclusively on the basis of profiling: when the function of the system is to support the human decision and not to replace it. Going back to the terminology of Art. 22.1 of the GDPR, it seems that what the AIA is proscribing are fully automated risk assessment decisions, whereas in principle partially automated decisions are admissible, subject to what will be said below.

3. "High-risk" predictive policing and its implications

In Annex III, paragraph 6(d) 'AI systems intended to be used by law enforcement authorities or on their behalf or by Union institutions, bodies, offices or agencies in support of law enforcement authorities for assessing the risk of a natural person offending or re-offending not solely on the basis of the profiling of natural persons as referred to in Article 3(4) of Directive (EU) 2016/680, or to assess personality traits and characteristics or past criminal behaviour of natural persons or groups'. The substantial difference between the factual scenarios referred to in the rule is that, in the case of high-risk use, the assessment is 'not based solely' on profiling. As can be seen, this

[60] Article 29 Data Protection Working Party, op.cit, p. 23.

[61] In detail Lazcoz, G., de Hert, P., "Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities", *Computer Law & Security Review*, Volume 50, (2023).

practice is configured as an attenuated modality in terms of risk of prohibited use: profiling is present but is not determinant in the assessment. This does not mean that the system is no longer influential (think of automation bias) and therefore its shortcomings still present risks that need to be mitigated.

Here we can see an evolution from the Manichean position of the Parliament, in which PersonBPP risks were not managed but prohibited. The final text establishes a response proportional to the risk. As mentioned above, the presence of a high-risk use with very similar characteristics gives substance to the factual assumption of prohibited use: the only uses excluded are the ones in which the assessment is not carried out exclusively by automated means.

The question, *a priori* not a simple one, will be to clarify in each specific case whether the implementation of an AI system falls into one or the other category. In the terms advocated, if the key element is the role played by the AI in the risk assessment, whether it is auxiliary or decisive, the *praxis* will have a significant influence on the legal qualification. It will be up to the producer in the risk assessment (art. 9) to specifically contemplate that the implementers of the system use it in a different sense to the original one, operating a delegation of functions that, if it had been considered at the beginning, could have determined the qualification of this as a prohibited practice.

The rule also sets out other high-risk AI scenarios that may involve predictive policing activities. Firstly, 'AI systems intended to be used by or on behalf of law enforcement authorities, or by Union institutions, agencies, offices or bodies in support of law enforcement authorities or on their behalf to assess the risk of a natural person becoming the victim of criminal offences'. On the other hand, "AI systems intended to be used by a judicial authority or on their behalf to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts, or to be used in a similar way in alternative dispute resolution". This could logically affect risk assessment systems that do not fall under the aforementioned paragraph 6 d), sometimes applying to systems that are not strictly police but judicial (indeed, we are thinking of predictive sentencing).

As mentioned above, with the AIA using such a broad concept as profiling, it is difficult for most uses of predictive policing, unless it is concluded that they are not AI, to escape the requirements of the regulation. The notable exception is place-based predictive tools. These were present in both the Commission's proposal and the Parliament's amendments, described as a high-risk use. On the contrary, the Council's common position moved them away from its version of Annex III by deleting the aforementioned paragraph 6(g): 'AI systems intended to be used to perform criminal offence analysis in relation to natural persons to enable law enforcement authorities to examine large sets of complex linked and unlinked data available in different sources or formats to detect unknown patterns or discover hidden relationships in the data'. The final text is aligned with that of the Council, and deletes paragraph 6(g).

Therefore, PlaceBPP tools, configured to detect where and when a crime is most likely to take place, are not considered high-risk systems under paragraph 6 d). According to this provision, the object of the assessment must be natural persons, a characteristic that, at least directly, is not present in these tools. However, it is true that the aforementioned paragraph 6 d) seems to mention two different uses of AI. On the one hand, "for assessing the risk of a natural person offending or re-offending" and on the other hand, "to assess personality traits and characteristics or past criminal behaviour of natural persons or groups". The question will be to determine to what extent a system that assesses the likelihood of criminal behaviour at a hot spot means assessing the traits and characteristics of the individuals or groups that transit those hot spots[62].

4. Conclusions

In general terms, the AIA provides a proportionate response to the risks posed by predictive policing systems. However, it uses concepts drawn from data protection law, such as profiling, to establish the dangerousness of these systems. This raises a number of problems.

The first is the exclusion of Place Based Predictive Policing tools. As we have previously pointed out, these systems also present ethical risks which, in the case of their poor design or use, could lead to the possible infringement of fundamental rights. And yet, they fall outside the scope of the Regulation and most likely also outside the scope of data protection law. This could leave European citizens unprotected against these systems, also taking into account the limits that will be imposed on Member States in what follows to regulate AI.

The second problem is that the criterion for determining the boundary between a prohibited practice and a high-risk practice appears to be heavily influenced by *praxis,* and cannot be determined at the time of system design. Indeed, what it is to adopt a risk prognosis on criminal dangerousness based

[62] The FRA objected to this possibility, as the predictions made by these systems do not include personal data but "aggregated statistics". FRA, *Bias in algorithms Artificial Intelligence and discrimination,* 2022, Available at: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf

"solely" on profiling, depends on the role of the legal operator in making the decision. Analysis of the context in which the system is embedded will not only help us to determine whether the system's decision is "formally" accepted as binding, but also whether, de facto, there is a normal course of action in which the system does not assist the AI deployer, but replaces its decision.

Finally, another relevant issue is that the AIA does not pay attention, at least indirectly, to the computational technique used to determine the level of risk (unacceptable risk or high risk). In other words, whether we are dealing with a classical actuarial technique, or a system based on machine learning, seems to be irrelevant. As we say, this is problematic, due to the additional ethical risks that systems falling into the second category carry out. It seems that, in this case, the AIA again follows too closely Art. 22 of the GDPR, whose factual assumption refers to automated decisions, but not necessarily to those taken by AI[63].

On the contrary, it could be argued that this element, although not introduced in the definition of prohibited practices, is deduced from the very definition of Artificial Intelligence adopted by the Regulation. In other words, only those systems with "inference" capabilities would be subject to the AIA, among which the recitals cite machine learning systems and those based on logic and knowledge approaches, excluding "traditional software based on rules defined only by humans and which automatically executes operations" (recital 12). Obviously, the relevance of this issue could be fundamental in resolving this interpretative problem, since following the restricted concept that some of the recitals seem to allude to, would significantly limit the tools covered and therefore the objects of prohibition and regulation as high risk. If we consider Spanish practice, there is no Person Based Predictive Policing system that has been trained through machine learning, which has not precluded either that these systems have been used and even institutionalised, or that they have not been critically analysed for the potential risks associated with their use[64]. So the legal operator would be faced with a Solomonic decision: either to disregard the computational technique used to assess the risk altogether or, on the contrary, to make it the most relevant parameter to the extent that it determines the overall subjection to the Regulation, or the total

[63] Palma Ortigosa, A., *Decisiones automatizadas y protección de datos. Especial atención a los sistemas de inteligencia artificial*, Dykinson, Madrid, (2022).

[64] See López Riba, JM., "Inteligencia artificial y control policial. Cuestiones para un debate criminológico frente al hype", in press, (2024), also Martínez Garay, L., "Evidence-based sentencing y evidencia científica", in Miró Llinares, F., and Fuentes Ossorio, J. L., *El Derecho penal ante lo empírico. Sobre el acercamiento del Derecho penal y la Política Criminal a la realidad empírica*, Marcial Pons, Madrid, (2022).

absence of guarantees in relation to a system. Obviously, it is not up to us to carry out an exhaustive analysis of what is or is not Artificial Intelligence, but we could venture to make three considerations that should be taken into account when it comes to informing the interpretation for its application to these cases.

First, that the AIA has to be interpreted taking into account all the interests at stake related to the use of these systems, and not only those related to innovation and the functioning of the market, but especially those related to the protection of fundamental rights. Second, that not only those systems that fall more clearly within the notion of AI may entail ethical risks and should be subject to evaluation. Thirdly, before regulating and interpreting the normative terms of what is regulated, it is essential to understand the consequences in practice. Opting for a broad conception can lead us to prohibit the use of systems whose ethical risks are in no way proportionally associated with what such a "sanction" would entail, and it would be illogical to prohibit some systems that may require supervision and control but which, *per se*, are no worse, in terms not only of effectiveness but also of traceability, transparency and guarantees, than classic ways of carrying out police activity. But using an overly restrictive conception of what is AI might neglect the risks associated with technologies that are not as computationally developed but equally or more dangerous in other respects. If such a conception is followed, then we should be able to find other ways of making normative demands on these other predictive policing systems that, based on the risk associated with the use of such systems, are also being demanded of those using machine learning.

The Regulation has made significant progress in the regulation of predictive policing, but there are still interpretative questions that will determine in practice what is or is not prohibited in the sector. In order to resolve these questions, we cannot rely solely on previous logics, such as data protection. Empirical knowledge about existing risks should also inform our interpretation of what is and is not prohibited in the framework of policing, and why not also: the "nebulous" concept of AI.

THE APPLICABILITY OF THE ARTIFICIAL INTELLIGENCE ACT TO THE HEALTH SECTOR AND SPECIALITIES REGARDING ITS COMPLIANCE

Iñigo De Miguel Beriain

Ikerbasque research professor. Researcher University of the Basque Country/Euskal Herriko Unibertsitatea. Member of the Spanish Bioethics Committee

I. Introduction

The approval of the AIA[1] is a particularly important event in terms of the regulation of this type of system in the EU, as it will introduce legal certainty in areas in which, until now, there were hardly any references in this respect. However, it should be remembered that the AIA has been designed with the aim of constituting a basic regulation which, in certain cases, must be interpreted in accordance with the regulation of specific sectors. This clearly applies to AI systems used for human health-related purposes, to which the European legislation on medical devices, explicitly referenced in the AIA, is generally applicable. These include at least the Medical Devices Regulation 2017/745 (MDR) and the In Vitro Diagnostic Medical Devices Regulation 2017/746 (IVDR), both of which are considerably complex.

In this chapter we will analyse the legal framework that will regulate AI tools for health purposes, and more particularly their qualification as high-risk systems on the basis of the provisions of the regulations just mentioned. In any case, we hope, at least, to be able to offer a precise description of the classification of AI tools used in human health in accordance with the risk-based scheme implemented by the AIA, as well as the possible difficulties that may arise from the differences in approach between this regulation and those of the medical devices mentioned above. To this end, we will begin by first outlining the relevant provisions of the AIA.

[1] The author wishes to acknowledge the support received from the Basque Government through grants to research groups (GI CISJANT, ref. IT1541-22). This work is part of the GODAS project (Project PID2022-137140OB-I00 funded by the MCIN/AEI/10.13039/501100011033/FEDER, EU) and the agreement between the public entity RED.ES and a consortium of six entities, including the UPV/EHU, to promote the implementation of the charter of digital rights in specific environments.

II. Regulation of medical devices in the AI Act and their consideration as high-risk by Annex I or III

1. Preliminary analysis: the regulation of medical devices incorporating Artificial Intelligence in the AI Act

As has already been mentioned in other chapters of this book, the AIA is based on the concept of risk: the degree of risk involved in the use of a system will determine the fundamental aspects of its legal status. Among other things, it will dictate a key issue: the requirements that the various actors involved in the system (providers, distributors, importers, etc.) will have to fulfil before and after its introduction into healthcare practice, as well as the supervision process associated with its approval. Therefore, the essential problem to be elucidated in this text is how to decide whether or not an AI that will be associated with healthcare purposes constitutes a high-risk system.

The answer to this question is to be found in Article 6 of the AIA, which specifies the classification rules for AI systems: they shall be considered as high risk when, by virtue of their characteristics, they are likely to be covered by the description in Annex III of the standard or when they fulfil two conditions: they are intended to be used as a safety component of one of the products covered by the Union harmonisation legislation listed in Annex I of the AIA, or they are themselves one of these products and must be subject to a conformity assessment carried out by an independent body for their placing on the market or putting into service, in accordance with the sectoral regulations. In the following sections we will analyse both routes separately.

2. Systems which are high risk in accordance with the provisions of Annex III

We have just explained that there are two main ways in which a system can be classified as high risk in the case of those used for health purposes. Let us begin by looking at the systems listed in Annex III, point 5, of the Regulation, which are those that the regulation describes without referring to health regulations. According to the Commission's original proposal, there would be two main types of high-risk products. Firstly, 'AI systems intended to be used for the assessment and classification of emergency calls made by natural persons or for dispatching or prioritisation of dispatching first responder services in emergency situations, for example, police, fire and medical services, and in patient triage systems in the context of emergency healthcare' (Annex III, point 6(d)). Recital 58 of the consolidated text explains well the reason for this decision, stating that, "AI systems used to evaluate and classify emergency calls by natural persons or to dispatch or establish priority in the dispatching of emergency first response services, including by police, firefighters and medical aid, as well as of emergency healthcare patient triage systems, should also be classified as high-risk since they make decisions in very critical situations for the life and health of persons and their property". To this first type of system should be added another: "AI systems intended to be used by public authorities or on behalf of public authorities to evaluate the eligibility of natural persons for essential public assistance benefits and services, including healthcare services, as well as to grant, reduce, revoke, or reclaim such benefits and services" (Annex III, in this case, point 5 a).

The Commission's initial stance was subject to alternative responses in the Parliament and Council versions. In particular, the Parliament opted to narrow the range of affected systems, limiting it to systems posing significant risk or harm to the health, safety or fundamental rights of individuals, while, on the other hand, advocating the inclusion in the high-risk category of systems entailing significant risk or harm to the environment. The Council, for its part, proposed an alternative wording to Article 6.2, which omitted any reference to Annex III. Both positions were rejected during the negotiations, leaving the Commission's original proposal intact, although the Parliament's nuances were to some extent reflected in the exception to the general regime for the systems described in Annex III, which we will analyse later.

However, where changes were introduced during the processing of the regulation was in Annex III itself. In contrast to the Commission's original wording, the Parliament wanted to introduce an additional letter (ba). This would mean that AI systems intended to be used to make decisions or materially influence the eligibility of natural persons for health and life insurance would be considered high risk. The final version does not take up this proposal, but rather a reasonably similar one presented by the Council, according to which AI systems intended to be used for risk assessment and pricing in relation to natural persons and in the case of life and health insurance (point c bis) would be high risk systems[2]. The reason for the inclusion of these systems in the high risk category is given in Recital 58 of the final version: "AI systems intended to be used for risk assessment and pricing in relation to natural persons for health and life insurance can also have a significant impact on persons' livelihood and if not duly designed, developed and used, can infringe their fundamental rights and can lead to

[2] It is perhaps worth noting that the Council's proposal included an exception to this general rule for systems developed for their own use by suppliers that were small businesses, which was not successful.

serious consequences for people's life and health, including financial exclusion and discrimination".

In turn, and as regards the qualification of AI systems intended to be used by public authorities or by a third party on their behalf to assess the eligibility of natural persons to access public assistance benefits and services, as well as to grant, reduce, withdraw or recover such benefits and services as high risk, the final version adopted contains an alternative wording with an important novelty: the range of high-risk systems is reduced to those associated with *essential* public assistance benefits and services*,* which is, in our opinion, a sensible alternative, present in both Parliament's and the Council's version, which will prevent certain systems that do not excessively alter the goods and rights they are intended to protect from having to be subject to the requirements of high-risk systems. On the other hand, the initial proposal has been improved by explicitly stating that public assistance benefits and services include health care, thus clearing up any possible confusion in this respect.

Finally, with regard to the use of AI systems in the case of emergency situations, there are two important novelties in the final text of the document, compared to what was evident in the Commission's proposal. Firstly, the material scope of systems classified as high-risk is extended to include those designed to assess and classify emergency calls from natural persons. This incorporates into the text an amendment from Parliament that is clearly aimed at preventing tools that can take vital decisions from having an adequate supervisory system. The second novelty is that the provision is also extended to patient triage systems. This provision, again introduced by the Parliament, addresses the proposals made by some authors to introduce AI systems in emergency triage[3]. In our view, the substance of the issue is more than reasonable, although it could be objected that it was probably not necessary to introduce such a tagline, as it seems obvious that such systems already clearly belong to the category of those to be used to establish priority in the provision of care and medicines.

3. Systems which may or may not be high risk as set defined in Annex I.

The second way to consider an AI system to be high risk is to include an explicit reference to the Union harmonisation legislation listed in Annex I of

[3] de Miguel Beriain, I. *The Ethical, Legal and Social Issues of Pandemics: An Analysis from the EU Perspective*. Springer, 2022; Weisberg EM, Chu LC, Fishman EK. The first use of Artificial Intelligence (AI) in the ER: triage not diagnosis. Emerg Radiol. 2020 Aug;27(4):361-366; Townsend BA, Plant KL, Hodge VJ, Ashaolu O, Calinescu R. Medical practitioner perspectives on AI in emergency triage. Front Digit Health. 2023 Dec 6;5:1297073.

the Act. The Council wanted to simplify the rule by deleting the reference to the Annex and instead including a sentence referring to the requirement to undergo a declaration of conformity as the key to deciding on the level of risk. This amendment does not appear in the final text, which basically takes over the Commission's proposal. It is therefore necessary to refer to Annex I, Section A of the Regulation to determine which AI systems applied to health will be high risk. Point 11 of the Annex includes an express reference to "Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (OJ L 117, 5.5.2017, p. 1)" (MDR). Point 12 of the same Annex I cites Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU (IVDR).

It should be borne in mind that this reference is crucial, since it is in these Regulations that we will find the answer to the question of whether an AI system will have to undergo a conformity assessment carried out by an independent body for its placing on the market or putting into service, which is the really crucial question for defining the qualification of the system. Hence, in order to establish the legal status of an AI system, it is necessary to set out the regime created by the MDR and the IVDR, which is precisely the task we will address in the following section.

However, before entering into this analysis, it is necessary to address another preliminary analysis: that of which AI systems are to be considered as medical devices and which are not, since only if an AI system is indeed a medical device, or constitutes a safety component of such a device, will it make sense to determine whether it has to undergo a conformity assessment by an independent body in accordance with health regulations. On the other hand, if we consider that it is not a medical device, the analysis of the risk inherent in the AI tool will have to be carried out in other ways, which are now outside the scope of this chapter. Having explained the significance of this particular point, we will immediately proceed to explain the conceptual framework of the MDR and the IVDR.

III. The regulation of medical devices: the MDR and IVDR provisions

1. Medical devices. A characterisation

According to Article 2.1 of the MDR, a medical device is any instrument, device, hardware, software, implant, reagent, material or other article intended by the manufacturer to be applied to human beings, separately or in combination with a medical device, which does not exert its principal intended action inside or on the surface of the human body by pharmacological, immunological or metabolic mechanisms (but to whose function such mechanisms may contribute), and which is used for specific medical purposes[4]. First and foremost, therefore, we are –among other things– dealing with a computer programme or similar. In turn, the IVDR qualifies as an in vitro diagnostic medical device "any medical device" which is a reagent, reagent product, calibrator, control material, kit, instrument, apparatus, piece of equipment, software or system, whether used alone or in combination, intended by the manufacturer to be used in vitro for the examination of specimens, including blood and tissue donations, derived from the human body, solely or principally for the purpose of providing information on one or more of the following:

(a) concerning a physiological or pathological process or state;

(b) concerning congenital physical or mental impairments;

(c) concerning the predisposition to a medical condition or a disease;

(d) to determine the safety and compatibility with potential recipients;

(e) to predict treatment response or reactions;

(f) to define or onitoring therapeutic measures.

In addition, the IVDR clarifies that specimen receptacles should also be considered as such products.

Putting the definitions of both regulations together, we have, in short, a fairly broad catalogue of what medical devices are. However, a reading of the MDR leaves at least two conceptual issues unresolved. The first concerns the characterisation of AI systems as software. The second concerns the notion of "specific medical purposes".

As regards the first, it should be recalled that AI systems are, according to Article 3.1 of the AIA, software, it is obvious that, if they are used for the purposes just specified, they are to be considered as medical devices and are therefore subject to the provisions of the MDR[5]. Therefore, this rule will apply to AI systems irrespective of whether they are an executable program, an

[4] Among those described in the article itself are the following:

- diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of a disease,

- diagnosis, follow-up, treatment, relief or compensation for an injury or disability,

- investigation, replacement or modification of the anatomy or of a physiological or pathological process or state,

- obtaining information by in vitro examination of samples from the human body, including organ, blood and tissue donations.

[5] Kiseleva, A. (2020). AI as a medical device: is it enough to ensure performance transpar-

interactive website, a web service, a script or a simple macro in a spreadsheet. Furthermore, the qualification as a medical device will be independent of whether the processing is simple or complex, the risk posed by the software to the patient or user, whether it is used by a healthcare professional or a profane, and the computing platform on which it operates, as long as its use with human beings or their data is intended for medical purposes[6].

Having clarified this first point, let us focus on the second: what exactly does "specific medical purpose" mean? Here there is a slightly larger gap and uncertainty, as not every AI system used in the field of healthcare is considered to be a medical device. This was already pointed out by the CJEU in a judgment concerning Directive 93/42, now repealed: "*The legislator has therefore made it clear that, in relation to software, in order for it to fall within the scope of Directive 93/42, it is not sufficient that it is used in a healthcare context, but it is necessary that its purpose, as defined by its manufacturer, is specifically medical*"[7]. Following this approach, the Medical Devices Coordination Group (MDGC)[8] has interpreted that software that is limited to storage, archiving, communication, or search tasks should not be considered as a medical device if it does not have a medical purpose. This includes, for example, software dedicated to altering the representation of data to improve the quality of its presentation or its compatibility. Nor, of course, should tools used to generate invoices or organise healthcare workers. On the other hand, a programme that searches an image for findings that support a clinical hypothesis in terms of diagnosis or therapy progression, or that locally magnifies the contrast of the finding on an image display to support a decision or suggest an action to be taken by the user[9] should be considered as a medical device. So are devices for monitoring or

ency and accountability? *EPLR*, *4*, 5.

[6] Beckers, R., Kwade, Z., & Zanca, F. (2021). The EU medical device regulation: Implications for Artificial Intelligence-based medical device software in medical physics. *Physica Medica*, *83*, 1-8.

[7] Judgment of the Court (Third Chamber) of 22 November 2012, *Brain Products GmbH v BioSemi VOF and Others*, Case C-219/11, par. 17.

[8] The Medical Devices Coordination Group (MDCG) was established by Article 103 of the MDR. The MDCG is composed of representatives of all Member States and is chaired by a representative of the European Commission. Its documents are not European Commission documents and cannot be considered as reflecting the official position of the European Commission. Nor are they legally binding (only the Court of Justice of the European Union can give binding interpretations of EU law), but they can serve as a basis on which to build an approximation of the concepts contained in the Regulation that created the Group.

[9] Medical Device Coordination Group (MDGC), "Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR (MDCG 2019-11)", in:

supporting conception or devices intended specifically for the cleaning, disinfection or sterilisation of devices used for the medical purposes described in Article 2.1 of the MDR or listed in Annex XVI of the MDR[10].

Finally, it is important to note that we are only dealing with a health product if its purpose is to benefit individual patients. If, on the other hand, we are dealing with programmes that are intended only to aggregate population data, provide generic diagnostic pathways or generic treatment (not directed at individual patients), improve scientific literature, medical atlases, or models and templates, or software intended only for epidemiological studies or registries, they will not be health devices and will therefore fall outside the MDR framework.

2. Classes of medical devices and supervision requirements inherent to each type according to the MDR

Having outlined the general criteria for when to consider an AI system as a medical device, it is now time to focus on the type of device involved, which will have important consequences for the approval and oversight process of the device. In this section, we will focus on medical devices, leaving the discussion of in vitro diagnostics for the next section. The medical device regulations state that medical software can be classified into several different classes: I, IIa, IIb and III. Rule 11 of Annex VIII of the MDR, which states the following, is to be applied to determine the class to which software independent of any other product belongs:

Software intended to provide information which is used to take decisions with diagnosis or therapeutic purposes is classified as class IIa, except if such decisions have an impact that may cause:

- death or an irreversible deterioration of a person's state of health, in which case it is in Class III; or

- a serious deterioration of a person's state of health or a surgical intervention, in which case they are classified as class IIb

Software intended to monitor physiological processes is classified as class IIa, except if it is intended for monitoring of vital physiological parameters, where the nature of variations of those parameters is such that it could result in immediate danger to the patient, in which case it is classified as class Iib.

All other software is classified as class I.

Following this rule, software used to calculate doses of highly toxic drugs, suggest a diagnosis or assist in planning therapies, or radiation, would be class

[10] See Article 2.1 of the MDR.

III, as an error could cause death. If an error is highly unlikely to cause death, it could be class IIb, while only those where an error cannot be expected to cause a serious deterioration of a person's health status can be class IIa[11]. If the system has several possible uses, its most critical specified use shall be considered for classification[12].

The MDCG Guidance on Rating and Classification of Health Software, however, seems to temper this framework. For example, it suggests that software intended to classify therapeutic suggestions for a healthcare professional based on patient history, imaging test results, and patient characteristics should be classified as class IIa, although it could be interpreted as class III, as an error could lead to patient death. Some other examples that may be useful in interpreting the system established by the MDR[13] are as follows:

- A computer programme intended to make diagnoses using image analysis to make treatment decisions in patients with acute stroke should be classified as class III under Rule 11(a).

- A diagnostic computer programme intended to score depression based on data entered about a patient's symptoms (e.g. mood, anxiety) should be classified as class IIb under Rule 11(a).

- A computer programme intended to rank therapeutic suggestions for a healthcare professional based on patient history, imaging test results and patient characteristics, for example, that lists and ranks all available chemotherapy options for BRCA-positive individuals should be classified as class IIa under Rule 11(a).

- An app aims to aid conception by calculating the user's fertility status based on a validated statistical algorithm. The user enters health data, such as basal body temperature (BBT) and days of menstruation, to track and predict ovulation. The fertility status of the current day is reflected in one of three indicator lights: red (fertile), green (infertile) or yellow (fluctuating phase of the cycle). This application should be classified as class I according to Rule 11(c).

The type of qualification that an AI tool obtains within this classification, as we have said, will determine, in accordance with the MDR scheme, the type of clinical evaluation required for product certification (CE) or the

[11] Keutzer, L., & Simonsson, U. S. (2020). Medical device apps: an introduction to regulatory affairs for developers. *JMIR mHealth and uHealth*, *8*(6), e17567.

[12] AEMPS, Guidance For Manufacturers Of Class I Medical Devices December 2019, revised July 2020, page 13, at: https://www.aemps.gob.es/productosSanitarios/docs/guia_fabricantes-ps.pdf

[13] The translation is by Guillermo Lazcoz Moratinos. It can be found in: Lazcoz Moratinos, G. (2023). Governance and Human Oversight of Automated Decision Making Based on Profiling, PhD thesis, available at: https://addi.ehu.es/handle/10810/61322

post-market surveillance to which it must be subjected[14], as well as what obliges the intervention of a third party in the process. As explained in Recital 60 of the MDR, "*conformity assessment procedure for class I devices should be carried out, as a general rule, under the sole responsibility of manufacturers in view of the low level of vulnerability associated with such devices. For class IIa, class IIb and class III devices, an appropriate level of involvement of a notified body should be compulsory*".

However, we are now exclusively concerned with whether or not such third party intervention is mandatory, which is the case for IIa, IIb and III systems. Also, by the way, even in class I, if such devices are placed on the market in sterile conditions, have measuring functions or are reusable surgical instruments[15], although the involvement of the notified body in such cases will be limited to verifying very specific aspects of such devices. In this respect, the MDR substantially changed the framework outlined by the previous Directive[16], which dictated that most stand-alone software, including applications, should be classified as class l or not designated as medical devices at all[17]. Finally, it should be underlined that the Medical Devices Coordination Group Guidance[18] clarifies that in case of any change in both the intended purpose and the clinical care context/situation in which the same device is used, the qualification could be altered, replacing the current qualification by a different risk class.

3. IVDR and in vitro diagnostic devices

What about in vitro diagnostic devices? Here it is necessary to refer to the IVDR, which uses a system relatively similar to that of the MDR, except that in this case the devices are divided into four classes, A, B, C and D, which are established taking into account the intended purpose of the products and

[14] This is to be understood as "all activities carried out by manufacturers in cooperation with other economic operators to institute and keep up to date a systematic procedure to proactively collect and review experience gained from devices they place on the market, make available on the market or put into service for the purpose of identifying any need to immediately apply any necessary corrective or preventive actions" (Art. 2.60 MDR).

[15] See: AEMPS, GUIDE FOR MANUFACTURERS OF MEDICAL DEVICES CLASS I, December 2019 July 2020 rev.1, p. 6, at: https://www.aemps.gob.es/productosSanitarios/docs/guia_fabricantes-ps.pdf.

[16] Medical Device Directive (MDD) 93/42/EEC.

[17] Keutzer, L., & Simonsson, U. S. (2020). Medical device apps: an introduction to regulatory affairs for developers. *JMIR mHealth and uHealth*, *8*(6), e17567.

[18] Medical Device Coordination Group (MDGC), "Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR (MDCG 2019-11)".

their inherent risks. The classification will be carried out in accordance with Annex VIII of the Regulation, which includes seven rules for the qualification of devices, of a certain technical complexity. The general rule is that the order of risk is incremental, with class A being assigned to low-risk devices and class D to devices representing the highest risk. The application of the IVDR risk classification system requires the involvement of a notified body for the approval of all non-sterile devices except for class A devices. With this in mind, it is estimated that 90% of IVD devices will be subject to a notified body review, compared to the 15% that had to comply with this requirement under the previous directive[19].

4. Exceptions to the general regime for the systems included in Annex III

Based on the explanation provided in the previous section, it seems inevitable to conclude that there will be many medical devices and in vitro diagnostic devices that are likely to be included into one of the categories that require supervision by a notified body. This, bearing in mind that it is the mere fact of the intervention of a notified body that is essential in determining whether or not an AI system applied in the healthcare setting is high risk, would mean that many of these systems would be considered as high risk, as very few of them are likely to be classified as class I in the MDR scheme (and there may even be exceptions in class I, as we have indicated)[20] or class A in the IVDR scheme.

However, there is an exception to this general rule, due to the outcome of the negotiation between the three European institutions. The essential change in the final version of the AIA compared to the Commission's original proposal is that it provides for the possibility that some of the systems included in Annex III may avoid such qualification under two conditions: that the deployer is a public law body or a private operator providing public services; and that the condition now included in Article 6.3 is met: that the system does not pose a significant risk to the health, safety or fundamental rights of individuals, including that it does not materially influence the out-

[19] https://www.tuvsud.com/es-es/industrias/asistencia-sanitaria-productos-sanitarios/diagnostico-in-vitro/aprobacion-certificacion-mercado/reglamento-ue-productos-sanitarios-diagnostico-in-vitro

BSI: IVDR Conformity Assessment Routes Notified Body Assessments, at: https://www.bsigroup.com/globalassets/meddev/localfiles/en-gb/documents/bsi-md-ivdr-conformity-assessment-routes-booklet-uk-en.pdf

[20] Grzybowski, A., & Brona, P. (2023). Approval and Certification of Ophthalmic AI Devices in the European Union. *Ophthalmology and Therapy*, *12*(2), 633-638.

come of decision-making[21]. This, the article clarifies, will be the case if any of the following conditions is fulfilled:

That the AI system is intended to perform a limited procedural task;

- the AI system is intended to perform a narrow procedural task;

- the AI system is intended to improve the result of a previously completed human activity;

- the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review;

- the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III.

In accordance with Article 6.4, it is for the provider to determine whether or not any of these circumstances apply. If, as a result of his assessment, he considers that an AI system referred to in Annex III is not high risk, he shall document his assessment before such a system is placed on the market or put into service. At the request of the competent national authorities, it shall provide them with the documentation of the assessment.

Obviously, the chosen resource to avoid high-risk classification –the inclusion of these specific criteria– has disadvantages that are easy to sense: technology may change substantially and the criteria set out may not respond to current needs; practice may reveal that there may be others that need to be added to the list; it may be complex to understand what realities they comprise in practice, etc. Hence, a number of provisions have been included in the final articles of the AIA to address these issues. Thus, firstly, according to Article 6.6, the Commission shall be empowered to adopt delegated acts in accordance with Article 97 of the Regulation to modify the criteria described above, either by adding new ones or by modifying existing ones, provided that there is concrete and reliable evidence of the existence of AI systems that fall within the scope of Annex III but do not pose a significant risk of harm to health, safety and fundamental rights. It may also, by means of delegated acts, waive any of the criteria laid down in Article 6.3 where there is concrete and reliable evidence that this is necessary to maintain the level of protection of health, safety and fundamental rights in the Union. In any case, it shall be nec-

[21] This result adequately responds to the intention, expressed in Recital 29, to limit the qualification of high risk to those AI systems identified as actually having a detrimental impact on the health, safety and fundamental rights of individuals in the Union and to minimise such limitation.

essary to ensure that the amendments will not cause a decrease in the overall level of protection of health, safety and fundamental rights in the Union.

On the other hand, Article 6.5 includes the provision that the Commission, after consulting the European AI Board, and no later than 2 February 2026, shall provide guidelines specifying the practical implementation of Article 6, including a comprehensive list of practical examples of high and non-high risk use cases in AI systems, in accordance with Article 96.

It should also be noted that Article 6 imposes a number of obligations on the provider. Thus, if he considers that an AI system referred to in Annex III is not high risk, he must document his assessment before such a system is placed on the market or put into service. It will also be subject to the registration requirement laid down in Article 49.2 of the Act. Upon request of the competent national authorities, the provider shall provide the documentation of the assessment.

IV. Recapitulation

The best way to summarise all that has been said in this chapter would probably be to emphasise that there are two main groups of AI systems for medical use that should be considered high risk: those that are likely to be used for some of the purposes described in Annex III of the Regulation without the provider being able to invoke any of the circumstances described in Article 6.3 to avoid the "high risk" classification, or those that, being in vitro medicinal products, medical devices or medical devices, require a third party conformity assessment.

In short, the system seems demanding in terms of risk assessment and even dysfunctional, given the conceptual differences between the applicable standards. There may in fact be a contrast between the risk measurement of the AIA and that of the health regulations. It may happen that a tool is assessed by a notified body under the MDR and IVDR as having a medium risk level (for example, class IIa or B) and yet it is considered "high risk" by the AIA, since, in the case of the latter standard, the requirement to obtain such a rating is precisely that it must be subject to such supervision. For this very reason, it makes perfect sense to have included in Recital 51 the idea that "the classification of an AI system as high-risk pursuant to this Regulation should not necessarily mean that the product whose safety component is the AI system, or the AI system itself as a product, is considered to be high-risk under the criteria established in the relevant Union harmonisation legislation that applies to the product. This is, in particular, the case for Regulations (EU)

2017/745 and (EU) 2017/746, where a third-party conformity assessment is provided for medium-risk and high-risk products." The meaning of "high risk" may or may not coincide in the case of one standard or another, so an AI system may be high risk in the terms of the AIA and not in the terms of the MDR, for example.

THE APPLICABILITY OF THE ARTIFICIAL INTELLIGENCE ACT TO THE FIELD OF PUBLIC ADMINISTRATION AND PUBLIC SERVICES AND SPECIAL FEATURES REGARDING COMPLIANCE: SPECIAL ATTENTION TO ANNEX III AND ADMINISTRATIVE ACTION AND PARTICULARITIES OF COMPLIANCE

Gal-la Barrachina Navarro and Andrés Boix Palop
Universitat de València and Senior Lecturer in Administrative Law Universitat de València

I. Introduction: the approach of the AI Act and the projection of its controls and guarantees on the actions of the public authorities

1. Overview: basic orientation and application to the action of public authorities of the Act

As is well known, and as this work has developed and explained more fully in other parts of this commentary, the AIA is not intended or defined in legal terms, in a specific manner, to be applied to the actions of public authorities (including the judiciary) in general, nor, more specifically, of public administrations, nor of the European Union itself or its Member States. It is a regulation that, applying the lessons learned from decades of public control over safety and control requirements with regard to the placing on the market of products (or, although less frequently, the provision of services that may also entail environmental or safety problems), establishes a series of protocols and requirements typical of this field. Thus, very quickly, together with the establishment of a series of uses of AI that are considered prohibited in any case in terms of services or products that could be deployed for certain purposes (Chapter II, Prohibited AI practices; Art. 5 AIA), and which in any case always have certain exceptions (and which, although at this point we refer to the commentary on prohibited policies, it should be noted that they imply limitations for both the public and private sectors, as we will see later, although the logic is still not so much to focus on the public sector as on the risks intrinsic to certain uses of AI), the rule goes on to delimit those uses that will result in greater regulation and compliance requirements and therefore in greater legal control (art. 6 and Annex III AIA), uses that in no case are defined in relation to the use that public authorities may make of AI systems, as this is not the focus, as has been said, of the AIA.

However, as we will see later, there are some uses that, of course, have a full impact on the sphere of action and possible uses of AI that public authorities may make, and these uses will therefore also be regulated by the Act, deploying obligations that will also have an impact on public authorities. A bit like what has also happened with data protection regulations (GDPR and the corresponding transposition regulations in each European country), which, although they are not rules specifically designed to regulate the actions of public authorities, but rather the economic operators and agents and subjects that act in the market, they have also ended up disciplining public authorities .

Furthermore, in the process of specifying and polishing the legal text, a whole series of uses of AI have been introduced in Annex III that are considered high-risk and have a more direct impact on the activity of public authorities (as we will see later, it can be considered that at present any action by any public authority, administrative, or judicial, that relies on the use of AI for decision-making, or assisting, or conditioning it, which impacts on the sphere of rights and obligations of citizens, by default, will always be considered as high-risk, even if this is not the essential objective of the rule, but merely an indirect, and fortunate, consequence of the AIA's desire to establish controls for private operators and market uses of AI).

The most important model of regulation and compliance (and the one that has the greatest impact on the effective deployment of AI tools today and in the future) is found in Chapter III AIA on AI systems defined as high risk, where, after the aforementioned delimitation (art. 6 and Annex III AIA), Section Two goes on to establish a series of technical requirements to be met (risk management, data management, documentation, information and transparency, robustness and security of the systems). As can be seen, none of these requirements, once again, is designed for public authorities, but they will also have to be complied with by the latter when they use AI for the exercise of public functions. Significantly, Section Three of this Chapter Two defines the obligations of the different providers and deployers of these systems, for the purpose of guaranteeing compliance with the legal demands and requirements established by the AIA in order to be able to place on the market products or provide services that integrate the use of AI, and does so by trying to cover the entire value chain so that there is always a person responsible for placing on the market or providing the service at European level who can be held liable for possible damages and breaches, as well as for the purpose of defining the specific obligations (and each other) of each of these agents.

In this sense, once again, it is easy to project these rules onto the public authorities in their various roles. However, in public action, it is uncommon for AI tools, public products, or services that incorporate AI to be placed on the market for others to market and integrate them into production chains. Instead, their position will typically be that of being responsible as the end user (who may also have defined and developed the use or use of the specific AI). In any case, it is not striking, as it is the usual trend, that the AIA includes a provision for this peculiar position of public authorities, which is worth repeating: public authorities will have the same obligations and requirements as a private agent in an equivalent position with respect to a specific AI system.

Once the system of requirements, demands, and obligations have been defined, Sections Four and Five of the AIA transfer to this sector a specific model of control to the systems of mandatory public technical standards and voluntary industrial standardisation (in a game explained in general terms by Álvarez García and which is also specified in Chapter X AIA with the regulation on codes of conduct) designed to guarantee industrial safety and consumer protection, as well as to facilitate the emergence of self-regulation rules and harmonisation systems in the markets, sometimes facilitated by the public authorities (regulated self-regulation), which is promoted on extensive previous experience. Thus, Section Four projects the typical control model for the field of AI, which is based on private entities that will essentially carry out the control, verification, and certification of compliance (notified bodies). These entities must, of course, always comply with a series of public requirements and controls that ensure their work is carried out correctly. These requirements and controls form the foundation of the compliance control model, guaranteeing its speed, efficiency in the market, and capacity for adaptation. Alongside these, there are obviously public notification authorities that must ensure that private agents comply with the regulations, verify that they comply correctly with their duties and that they have the technical capacity to do so, as well as intervene in cases of detection of serious non-compliance, in more classic administrative control and verification functions with respect to the actions of private agents (typical in any regulated market and a manifestation of the most basic administrative police).

On the basis of this framework, Section Five defines in what terms the conformity assessment, verified in principle by these notified bodies, of products and services incorporating AI must be carried out, which leads, as in any market, to obtaining certificates and labels (which are specified in the EC framework, art. 48 AIA) and their translation into registers for control purposes, which are what will allow the products to be placed on the market, but, significantly, this conformity assessment when they are systems used by public authorities will be carried out internally, without the need to resort to external controls. Once again, the whole system is defined from this traditional market perspective and without considering more demanding or different standards for public authorities and public administrations, given their particular position and their capacity to be able to harm citizens' rights to a greater extent. Moreover, when there are exceptions or singularities, these are in terms of deference (art. 111.2 AIA allows the entry into force of the requirements of the text to be deferred for up to six years when they are used by public authorities; for example, as we have pointed out, the conformity assessment for these cases is internal).

In any case, as is logical, these rules, in terms of their material content, must also be complied with by them when they use AI systems in the future to perform their functions or provide services. This is to make sure that the AI tools they use, which fall within the scope of application defined by art. 6 and Annex III (which we have already said will be practically all those that may be used by public authorities), regardless of any public, European or state controls that may be defined, must be subject to these controls with regard to their compliance, and incorporate the CE marking or equivalent for the public sector, as well as being subject to control and inspection for cases that may entail greater risks on the part of the administrative control authorities. As happens with any public administration when it makes use of a service or product placed on the market which has to comply with technical standards, and which, of course, can only be used or integrated in its actions and services if it has passed the corresponding controls in a satisfactory manner.

Chapters IV and V AIA, insofar as they incorporate specific obligations for certain specific AI uses, be they *chatbots* or general purpose AI models, do not project major issues on the public sector, beyond the fact that, if AI systems of these types are used by the public sector, these rules will have to be complied with, but there is no noteworthy specificity projected on the public sector. Chapters VI, VII and VIII AIA have a much greater impact on the public sector, insofar as they respectively establish measures for promotion (support for innovation in the sector), public governance (with the deployment of national and European control authorities in the sector), which have been reinforced as the negotiation of the legal text has progressed, and public registers which, of course, do form clear nuclei of public action. However, although in these cases there is clear administrative action on the sector, they are far removed from our object of interest, which is not so much how the Administrations and public authorities must act to promote, control, or guarantee the correct functioning of the AI rules on the AI market or the products and services that incorporate it, but rather with respect to the rules that the public authorities must comply with when they are the ones who use it. The same can be said of Chapter XI on the functioning of the committees or the organisational measures of the sanctioning power in Chapter XII.

It should be noted, however, that some measures in Chapter IX on post-market controls will also apply to the uses of AI by administrations and public authorities, as these will normally be uses that will continue over time. These obligations imply the need for controls on their use and implementation, requiring reporting of serious incidents and risk assessments (Art. 79 AIA) and specific notification of problems with regard to high-risk uses of AI (Art. 82 AIA) to which public authorities will of course be bound. Of course, with regard to non-compliance, Chapter XII establishes a sanctioning regime, once again designed for the market and companies rather than for public authorities (as the clearest example of this, sanctions are defined on the basis of the turnover of the company considered responsible), which will have to be adjusted, as has been done in the area of data protection, to the specificities of the public authorities.

Either way, and as we can see, we are dealing with a legal regime that has not been defined or developed with the public authorities in mind, so it will have to be complemented by the national control systems already existing in this area for the regulation of automated administrative actions or those that use algorithms or AI in each Member State, pending the existence of some harmonising European standard (experience in the field of data protection allows us to anticipate that if these come about they will not be very ambitious, in order to leave room for the administrative self-organisation of the internal public authorities), so that in this general characterisation of the functioning of these rules and in order to understand how they will be projected, we must briefly review the current control model of domestic law in the different Member States and also in Spain. In addition, and as indicated above, art. 111.2 AIA allows Member States to defer the applicability of the requirements and obligations set out in the rule to AI algorithms and solutions to be used by the public sector by up to 6 years, which gives the impression that, ultimately (especially given the foreseeable rapid evolution of the sector and its regulatory framework), converts all the rules detailed below into a kind of guidelines that Member States must subsequently decide whether or not to project on their public authorities, rather than into mandatory rules. In short, and as we have been saying, it is clear that we are not dealing with a regulation that is primarily intended to be applied to the AI solutions used by the public authorities, but which will only affect them indirectly and in a manner that is highly dependent on the way, form, and timing that the public authorities themselves consider to be most convenient for them.

2. Integration of the control of automated and algorithmic activity, and the use of Artificial Intelligence by public authorities with the rules of national law and some of their problems and weaknesses.

Both in our national system and in other Member States, public authorities can carry out part of their activity in an automated way and without human decision making, being possible to use AI (art. 41 of Law 40/2015, of 1 October, on the Legal Regime of the Public Sector (Ley 40/2015, de 1 de octubre, de Régimen Jurídico del Sector Público, RJSP). The use by public authorities of AI-based systems can bring obvious benefits, especially in terms of the effectiveness and efficiency of the system in its operation, or the adoption of discretionary decisions endowed with a certain impartiality and objectivity; however, such an advance entails risks and possible clashes with the fundamental rights, freedoms, and guarantees that the administration has a duty to protect. A regulation that provides legal certainty and ensures that the administration's guarantees will not be affected is essential, as optimisation of the system is desirable, but the protection of rights is unavoidable. Finding a regulation with a normative framework that allows progress to be made, but which at the same time ensures the legal status of the administered is a task that requires a delicate balance to be struck, which is not very easy to achieve when there are conflicting interests.

The substitution of human intelligence by Artificial Intelligence can be done in two ways: the decision adopted directly by the algorithm and without the intervention of human intelligence, and the one in which human intervention is involved, albeit in a secondary way, at the service of the AI. Within these two groups, there is doctrine that considers it essential to exclude the application of AI in certain cases, such as in discretionary decision-making, following the German solution while others, on the other hand, advocate its use in those decisions with a "low level of discretion or when the exercise of discretionary power involves the use of technical and not political criteria". This issue, regardless of the doctrinal discussions, has practical consequences, as the use of AI will be much more common and widespread if the second view is adopted. Moreover, the safeguards required by the two systems differ, since a system where AI can determine or assist in the delimitation of non-regulated decision making, which is where it will actually have a real differential utility, as this is where AI can bring about improvements (an algorithm automating regulated decisions is not really even a use of AI as defined in Art. 3 AIA), but in these cases, the possible risks for the rights of the administered persons are increased, not only because of problems of application or correction and efficiency, but also, for example, because of the very diverse and potentially serious effects on fundamental rights (Soriano Arnanz has analysed them in detail*; see* also his work of 2023). Notwithstanding the above, it is a mistake to consider that the mere final supervision of the natural person at the time of issuing the opinion or the administrative decision is a sufficient guarantee to trust that fundamental rights will not be affected, as the system itself has an inherent danger: its improvement generates overconfidence, and overconfidence undermines the perception of the need for supervision, even to the extent of conditioning one's own opinion, which will be at the expense of the AI's response. Solutions cannot always involve this human oversight, moreover, because in many cases it will be redundant or dysfunctional, although as we shall see this is an approach that the AIA has adopted for high-risk systems.

The first control mechanism that must guarantee the non-affectation of fundamental rights for the citizen, in contact with the AI decisions adopted in the public administration, must be a rule aware of the possible infringements that regulates the proper use, it will be the legal embodiment and projection on the sector of the precautionary principle, by which precautions must be taken, and the risk inherent in document management must be eliminated, for example, by isolating all data whose processing may generate infringement of rights. The administration must protect to a greater extent the rights affected by decision-making in its sphere with the application of the AI, but as we have already stated, applying the rule we are analysing to its full extent will generate a friction with the nature of the public sphere (which necessarily requires a certain laxity to protect the public system), ultimately detrimental to the rights of the citizen. The greatest dangers are generated, as is obvious, when the algorithm can become a substitute for law, which in no case occurs from a formal perspective if the system is endowed with a specific and guaranteeing legal regime, with adequate specific legislative coverage for the sphere of the administration, but can occur materially if the definition and specification of decisions with discretionary content that affect the sphere of rights and duties of citizens are not sufficiently outlined *ex ante* and a correct delimitation of the functioning of the models is not achieved in order to adapt them to the required public purposes. For this reason, one of us has extensively emphasised the need to understand, comprehend, order, and duly control the use of AI by the public sector that fulfils this materially regulatory effect through the introduction of additional guarantees and controls, a need that has not been addressed for the time being, and to which the AIA does not pay much attention either (due to this concern for a more market-oriented regulation of products and services). In these cases, AI will have materially normative effects (in the sphere of public administration, materially regulatory), since it is through this that the effective scope of action of the public authority will be specified in each case.

As we have already pointed out, none of this appears in the AIA, which is an instrument designed as a legal model of intervention that seeks to demand transparency in terms of access, significant control, external audits, and mechanisms specific to the private sector. In the future, in our opinion, we should aspire to be able to have our own regulation on the use of AI for public administrations, which differs from and goes beyond what it is established for private entities in all that is necessary to adapt the nature of the public sector and the protection of the fundamental rights of those with whom it works on a daily basis. A regulation that, within this framework, protects the citizen even more, to ensure that the process of substitution of decision-making by AI respects all the guarantees. This is something we consider necessary because the possibilities for public authorities to affect citizens' rights and their basic legal status, and specifically their capacity to harm them, are much greater than those of the private sector, something that we believe it is essential for the law to take into account. However, we are not at this point, nor is it the role of the AIA to do so, which is left to a later legislative stage and largely to the responsibility in domestic law of the Member States themselves, and even to the administrative self-organisation of each public authority. This is therefore a pending debate on which we shall say no more.

Having framed what the AIA aims to do and what it does not, and framed within this moderately critical analysis of the ambition of the current regulatory framework, we will focus the present study directly on the articles available in the AIA and we will carry out a descriptive analysis of the precepts that we consider, beyond the general control model described, most directly affect the public sector and public administration. Specifically, we will refer to:

- Recitals 4, 5, 6, 131 and 157.

- The prohibitions in Article 5, which are binding on administrations.

- The precautions of Article 6, in relation to the obligations imposed on public bodies by Article 27 (in connection with Articles 49 and 71).

- The situations referred to in Annex III which concern public authorities, in particular points 5 to 8.

- References to administrative actions in Articles 30, 34, 43, 45, 56, 57, 58, 59, 63, 66, 79, 82, 99 and 100.

II. Development, processing and final content of the precepts of the Act that most affect the public authorities

It is worth analysing, due to its interest, the evolution of the modifications and extension of the legislative ambition of the precepts that affect public authorities in the AIA, from the first proposal prepared by the European Commission in April 2021, to the latest text proposed by the Council, in the version that incorporates the amendments approved by the European Parliament, synthesising and specifying the part of the articles that really directly binds public administrations.

1. General considerations that can be projected singularly on public authorities in their use of Artificial Intelligence

The initial Proposal for the Act was based on an organisation of the regulation categorising the risk of the use of AI systems, from minimal or low risk to high or unacceptable risk, and this model is the one that was finally approved. For those uses considered to be of minimal risk, the AIA makes recommendations, which are accepted on a voluntary basis, with codes of conduct and good practices, and some obligation of transparency; on the other hand, systems of unacceptable risk are prohibited; at the intermediate point, and making up the bulk of the regulations, are those of high risk, which are of particular interest to us as they affect public sector affairs, although, as we shall see, with quite a few exceptions. The treatment in this regulatory scheme of the regulatory specificities affecting public authorities is sparse, and we can see this even in the reasoned explanatory memorandum itself of the need for regulation, where mentions of the differential position of public authorities are minor, sparse and with little structural impact on the regulation.

Thus, for example, Recital 4, analysing AI as a technological whole, recognising its rapid evolution and listing a series of benefits, although it focuses mainly on the economic and competitiveness aspect, adds references to social or environmental benefits and to some related areas, mentioning public services, security or justice. As can be seen, the focus is by no means on the public sector, but rather on the public sector as a tangential and indirect beneficiary of improvements and advances which are essentially taking place in other areas and which are spreading their benefits throughout the economy. However, Recital 5 has finally recognised certain risks to public interests and fundamental rights, categorising the damage as tangible or intangible, and the harm as physical, psychological, social, or economic. In this regard, we must remember once again that the impact on these rights, and therefore the inherent risks, are much greater when AI is used for administrative decision-making, due to the very nature of the public sector, and its duty of enhanced protection compared to that which applies to private entities. However, it does not appear that this view of the greater risks in these cases will be translated into regulation, as there will be no specific rules or additional safeguards. However, Recital 6, recognising the importance and impact of AI in several areas, and the need for the AIA to ensure respect for the values of the Union

(Art. 2 and 6 TEU, fundamental rights and freedoms of the Treaties and the Charter), allows these safeguards to be projected onto public activities.

Some more specificity on the public authority, as obligations can be drawn from it that must be taken into account very directly, derives from Recital 131, on transparency, when it establishes the duty of high-risk AI providers to whom Union harmonisation legislation would not in principle be imposed, and those who consider that they are not at high risk because an exception applies to them, to register in the EU database set up by the Commission. In particular, it imposes a duty on public authorities, bodies, offices, or agencies to register in the database, indicating the system they intend to use. Recital 157 refers to the scope of competence, establishing that its application is without prejudice to the competences, functions, powers, and independence of national public authorities or bodies, ensuring their access to the documentation created under the AIA. It adds in relation to the above, and especially with regard to the protection of fundamental rights, the need for a specific safeguard procedure to be applied when the AI presents a high risk to health, safety or fundamental rights: To high-risk systems, to prohibited systems placed on the market, to systems put into service or used in violation of the prohibited practices of the AIA and to systems placed on the market in violation of transparency requirements which present a risk.

2. Effect on public administrations and public authorities of the prohibited uses of Artificial Intelligence set out in Art. 5 AIA

Although the analysis of Art. 5 AIA and the prohibitions established therein are dealt with elsewhere in this work, it should be noted in this analysis of how the AIA affects the activity of public authorities that some of these prohibitions have a very direct impact on areas of public action which are now directly prohibited (in reality, they were already prohibited in all cases: the AIA simply means that certain actions which our public authorities could no longer carry out due to requirements derived from the basic protection of our fundamental rights cannot be carried out using AI either).

Art. 5 AIA tries to avoid this impact on citizens' rights by listing the AI practices that would be prohibited by the European Union, as they are potentially dangerous in terms of the violation of values and fundamental rights of the European order, or susceptible of exercising a general manipulation of the population and, above all, of the most unprotected and vulnerable sectors. It establishes a series of prohibited cases that have increased substantially from the European Commission's Proposal for a Regulation in April 2021 to the Common Position ("general approach") on the AI Law on 6 December

2022 of the European Council, with the Parliament's negotiating position on the AIA, June 2023 European Parliament, EPRS, finally grouping together a compendium with a series of common elements.

From this list of prohibitions, it makes little sense to focus on those that imply, as has been said, an absolute prohibition that could prevent public administrations from using AI for certain purposes or functions, but not because we want to prevent the use of AI for these activities, but because it is the activity itself that is prohibited as being incompatible with our system of guarantees, rights and the rule of law. For example, AI systems cannot be used to generate social credit models, simply because these models are understood to belong to dictatorships or authoritarian models, whether they are applied using AI or not. Similarly, constant surveillance of citizens is incompatible with a democratic model, whether using AI or not, and is therefore also prohibited when AI is used.

It is more interesting to note the cases in which certain generally prohibited uses of AI find a certain relaxation when they are carried out by public authorities for activities that are considered socially justified, exceptionally, because of their crime prevention purpose. Thus, we can highlight the perceived differentiation, within such high-risk systems that are normally totally prohibited, of remote biometric identification systems (unlike other systems, where some regulation is possible), which try to identify persons remotely with their biometric data incorporated in reference databases, requiring prior concession by a judicial authority or independent administration (Article 5.3) and subject to special transparency controls (Article 52) and systems that are in any case considered so risky that they are generally prohibited (Articles 5.1d) and 5.2). Thus, paragraph 5.1(d) prohibits the placing on the market, putting into service or use of AI to assess risks and likelihood of crime in natural persons, including profiling and personality assessment, the exception practically voids the prohibition for the public sphere, as it indicates that it shall not apply "*to AI systems used to support the human assessment of the involvement of a person in a criminal activity, which is already based on objective and verifiable facts directly linked to a criminal activity".* Paragraph h) specifies the use of such biometric identification "in real time" in public spaces, prohibiting its use, again except for the search for victims or missing persons, the prevention of threats to life or security (specifically noting terrorist attacks) or the location of suspects for criminal offences or the execution of sanctions for Annex II offences with sentences or security measures of at least 4 years. Paragraph 5.2 regulates certain aspects for assessing the application of point (h), in terms of confirming the identity of the person, the naturalness of the situation, the seriousness and extent of the non-use of the AI, and the impact on the rights and freedoms of the person concerned. The article refers to the national regulation of each member state, in terms of temporal, geographical and person-related limitations, and to the prior authorisation of the law enforcement authority, with an assessment of the impact on fundamental rights according to Article 27, which we will analyse later, and registered in the database system of Article 49. Again, the AIA plays with the counter-exception, even in this provision, by clarifying that "*however, in duly justified cases of urgency, the use of such systems may be commenced without the registration in the EU database, provided that such registration is completed without undue delay*".

The third point adds, as regards prior authorisation, by an independent judicial or administrative authority, the requirement of a reasoned request in accordance with national rules, and we again lower the safeguards, as it will not be necessary when the emergency situation is justified, provided it is subsequently requested without delay within a maximum of 24 hours, interrupted in case of refusal and the information deleted. A subsequent requirement is added for notification to the relevant market surveillance authority and national data protection authority, containing the information in paragraph 6, and "*without sensitive operational data*" with an obligation for them to communicate annual reports to the Commission (to be subsequently published by the Commission). The AIA calls on Member States to regulate in their national law the rules for the application, granting, exercise, supervision, and notification of authorisations, to notify the rules to the Commission (at the latest 30 days after their adoption), and to grant the option to regulate more strictly the use of biometric identification systems. In other words, the AIA seeks to establish an interventionist framework with minimum standards to be applied, which can be developed more restrictively, but not more laxly, by each national parliament.

As can be seen, in these cases the AIA lifts the absolute prohibition on the use of AI when it understands that there is a legitimate purpose, and in these cases the use becomes high-risk. It should be noted, however, that the lax and very broad definition of the assumptions that give way to the possibility of using these AI tools may allow a very generic appeal to these risks to enable a more massive use than is apparently intended by the rule (for example, the control of possible terrorist activities or the appeal to the prosecution of certain crimes may lead, and this has already been expressed by part of civil society) to support the activation of very generic controls, for example at borders, of AI systems that involve this type of functionalities. In order to prevent this exception from ending up having this effect, implementing legislation will be needed to limit and control this authorisation, both in European law and in its national integration.

3. Regarding the precautions of Art. 6 AIA on high-risk uses in relation to the resulting obligations for public authorities

Again, the analysis of art. 6 AIA, in conjunction with Annex III, for the purposes of the delimitation of the systems considered high risk by the AIA, from which a reinforced legal regime of requirements and demands is derived, corresponds to another part of this work. But let us try to map how this regulation is specifically projected with respect to public authorities.

Thus, in general, in order to be permitted the uses of AI considered as high risk by the conjunction of these precepts, certain requirements must be fulfilled throughout their existence. The specific precautions that will affect the public sphere, insofar as they are no different to those of any private subject, are therefore also found in art. 6 AIA itself, which must always be analysed in this case together with the additional obligations imposed in art. 27 AIA for distributors that are public law bodies or private entities that provide public services, which are directly and specifically designed for these cases. And all of this without forgetting that the general provisions relating to evaluations, certifications, and registers will always be essential in order to provide the necessary transparency and security to the AI system, within the high-risk cases in which the Administration is involved.

Beginning with the rules for classifying high-risk AI systems in art. 6 of the AIA, after the two generic conditions for classifying high risk, the provision adds the AI systems contemplated in Annex III. And here is one of the most important modifications throughout the legislative process: the cases not considered high risk are extended to include compliance with one of the following conditions (not applicable to the profiling of natural persons):

- Limited procedural tasks.
- Improvements to the result of a previously performed human activity.
- The detection of patterns in decision-making or the deviation of these patterns from previous decisions, without being intended to replace human assessment (without adequate review) or to influence it (something quite difficult to conceive in reality, as the AI tool tends by its very nature to end up absorbing the real and effective competence of the decision, once the human intelligence trusts it so much that it ceases to consider the need to intervene).
- Preparatory assessment tasks in Annex III uses.

As is well known, if the provider considers that its AI system complies with these requirements and therefore, despite being in Annex III, is not high risk, it must document its assessment prior to placing it on the market and register under Article 49.2 (the assessment must be documented at the request of the authorities, and the submission of such an assessment is therefore not even a prerequisite for placing it on the market). In addition, the Commission reserves the right to add new conditions or modify existing conditions (without reducing the overall level of protection) where it considers that systems covered by Annex III exist, but which, in its opinion, do not pose a significant risk of causing harm to the health, safety, or fundamental rights of natural persons.

In short, we must point out that in any case, administrative actions using AI are considered as explicitly indicated in Annex III, where AI in the public sector is generically included as high risk (although the option is subsequently opened to a certain degree of flexibility in some cases, with a series of fairly broad conditions that perhaps allow some acts to be extracted from this more protective category by considering them to have less impact on the rights and guarantees, and then adding the option of not applying the third paragraph, modifying or adding more exceptions, if in the Commission's opinion there is no effect on the rights or guarantees of the person administered we believe that this situation should be understood as exceptional with respect to uses that affect the sphere of rights and duties of citizens and that should only be applied to internal processes). And this is because the constant support in administrative tasks, just by judging the correctness of their decisions, with a tool that facilitates the work with such magnitude, inevitably generates a dependence from which it is very difficult to dissociate oneself, and although theoretically and formally it is the responsible person who issues the resolution, de facto it is the AI that produces it, replacing human intelligence, and finally leading to an independence of the algorithm to reach its own decisions. The lack of a concrete definition of the margins of what is to be considered high risk generates a more than patent danger for the rights of the person administered in a specific administrative process and, in general, for the rights of citizens.

With regard to Art. 27 AIA, concerning the impact on fundamental rights for high-risk AI, it is key to note that when those responsible for the deployment are public law bodies or private entities that provide public services, the AIA obliges them to assess the impact that the use of the AI may have on fundamental rights, always with particular rigour. This assessment (which must be carried out on the first use of the system, based on previous ones) consists, according to the articles of the AIA, of describing the processes of the deployer with their purpose, the period of time and frequency of use, the persons or groups likely to be affected by the system, the specific risks and human oversight measures, together with the measures to be taken if the risks materialise, along with mechanisms to reclaim. Following the assessment, the results must be notified to the market surveillance authority, unless exempted by the authority itself under Article 46.1 AIA.

Finally, with regard to the registration of the system as a guarantee and transparency mechanism, one of the main guarantees that must be in place, and which is regulated by art. 49 of the AIA in relation to art. 71 of the AIA, is that of publicity and registration control. With the use of the AI, control and transparency are necessary so that the authorities themselves and the citizens know where the use of the AI comes from, under what conditions the decision or resolution has been taken, with what considerations, and finally whether or not it has been reviewed by the functional manager who initializes it. If the AI predetermines an administrative decision, we must always make it possible for citizens to know that it exists and how it is being applied, so that citizens who are dissatisfied with the result and wish to appeal the corresponding act may have an interest in knowing the configuration of the algorithm, as well as its correct application in the specific case and it is essential that the AIA imposes on the Administration the duty to attend to such a claim. Art. 49 of the AIA establishes that prior to the introduction of the high-risk AI system on the market, it must be registered in the EU database. This database, according to Art. 71 AIA, shall be drawn up by the Commission in collaboration with the member states, in consultation with the relevant experts, divided into sections depending on the type and person subject to the obligation to register, accessible to the public as a means of guarantee, with the exception of sensitive data only visible to the market surveillance authorities and the Commission, which shall also be responsible for processing it, and shall provide support and accessibility to obliged and responsible parties. It should be noted that, if we find ourselves in the case of application of the exceptions in Article 6.3, the registration will be carried out by the provider or authorised representative, who will register the system himself, and also in the case of public authorities, bodies or agencies, they will be registered by their representatives with selection of the system and its use. The rule specifies that for high-risk AI systems in Annex III relating to the use of biometrics (point 1), law enforcement (point 6), and migration, asylum and border control management (point 7), the registration shall be carried out in a secure section, extending data protection with respect to the general registration with specific information to which only the Commission and national authorities shall have access.

4. Projection of the delimitation of high-risk uses according to Annex III on the use of Artificial Intelligence for public authorities

As we have already mentioned, the joint regulation of the control of the use of AI for private entities and for the public administration that the AIA has opted for, is problematic, as the competence of the public authorities affects and has a direct impact on the personal legal status of those affected in a much more sensitive way, due to the very nature of the relations of the public sector with those administered. The regulation raises certain doubts as to its application to Public Administrations, even from a systematic point of view, since its specific articles are more specifically included in this Annex III, the mere fact of regulating its specific application in an Annex shows the distance between legal situations, insofar as the basic structure of a framework designed for private entities and large companies may mean ignoring some additional risks for certain rights when applied to the public system. Restrictions, interventions, audits, and a series of self-protection conditions can be imposed on private actors that they must comply with, and which are sometimes difficult for the administration to transfer to their full extent (see in this respect the problems in the area of sanctions, similar to those that have already occurred in the area of data protection). But there are also additional problems affecting the public sector that this approach simply cannot address. There are obvious risks for the protection of citizens' rights when AI is used by the public sector, because of its greater capacity to affect. Nevertheless, the guarantees and safeguards that the AIA introduces for high-risk systems are at least a minimum on which to build, and which from now on will be required for all public uses that are deemed to fall within the definitions in Annex III.

In this sense, and on the basis of the legislative evolution of Annex III, from which certain minor inconsistencies can be deduced, since on occasions it seems that all uses of AI by the judiciary or public administrations are defined as high risk if they have to do with decision-making or the provision of public services, but on the other hand some additional provisions are added on specific services or specific actions that could raise doubts as to the greater intensity of control in some cases compared to others. In our opinion, the most correct and guaranteeing way of interpreting the AIA on this point is also the simplest: any use of AI to aid decision-making or directly replacing it completely with regard to the exercise of judicial or administrative functions or the provision of public services is currently considered, after the successive extensions made in the legislative process, to be high risk. And if there are additional provisions for more specific areas, this only reinforces such consideration with respect to these areas, obliging a more careful vision and a stricter application of the precautionary principle and to understand these cases to include even uses that, for example, indirectly affect decisions and actions.

In particular, Annex III mentions the high-risk AI systems under Article

6.2 in the following areas, focusing on those affecting the administration, which are those relating to points 5, 6, 7 and 8, to the specific wording of which reference should be made: access to and enjoyment of essential private services and essential public services and benefits: access to and enjoyment of essential private services and essential public services and benefits (5th); ensuring compliance with the law, insofar as their use is permitted by applicable Union or national law (6th); migration, asylum and border control management, insofar as their use is permitted by applicable Union or national law (7th); and administration of justice and democratic processes (8th).

5. Draft rules of the Administrative Proceedings Regulation

In addition to the above, the AIA contains references to administrative proceedings in Articles 30, 34, 43, 45, 56, 57, 58, 59, 63, 66, 79, 82, 99 and 100.

- We have already indicated that a registration procedure must be followed for high-risk AIs (even in some cases where they are exempted from such categorisation), and that this obligation is also incumbent on public administrations (Article 49 AIA in conjunction with Article 71):

The notifying authorities, which each state shall designate for the procedures of assessment, designation and notification of the conformity of the AI with the Regulation and its monitoring. The monitoring of compliance with the requirements imposed on high-risk AI systems will therefore be carried out by the notifying authority, a public body designated for this purpose by each Member State (Article 30), which will be responsible for determining the procedures for assessing compliance with the standard. These authorities will require documentation from the notified bodies, which in turn have certain operational obligations, as indicated in Article 34 (for AI tools used by private actors, since as we have pointed out for algorithms used by the public sector the conformity assessment is internal): They shall verify the compliance of high-risk AI systems with the assessment procedures of Article 43, avoiding unnecessary burdens, and considering the size of the provider, sector in which it operates, structure, degree of complexity of the AI, in order to minimise administrative burdens, while respecting the required stringency and level of protection. Notified bodies shall make available to the notifying authority, and submit on request, all documentation to enable the assessment, designation, notification and monitoring by the authority to be carried out. In Article 45 and following on from the above, the information obligations of notified bodies are mentioned.

Alongside this body, the AIA provides in Article 59 for the designation of the national supervisory authority (which may also exercise the function of market surveillance authority under Article 63) with the task of supervising the implementation and enforcement of the AIA, and representing its State in the European Artificial Intelligence Committee. This committee will ensure uniformity in the application of the AIA across the Member States. The national supervisory authority will be responsible for granting authorisation for the introduction or putting into service of high-risk AI on the market, and in accordance with Article 45, all such decisions must be subject to appeal.

- Special mention should be made of Articles 57 and 58 regarding the regulation of controlled test sites for AI, in this case the AIA no longer imposes a supervisory attitude on the Member State, but the creation of initial test sites, with support and advice from the Commission, where the states shall ensure the allocation of sufficient resources, and commit themselves to co-operation with the relevant authorities, so as to provide a safe environment that encourages innovation, testing and validation of innovative AI systems prior to market introduction and operation, with the authority providing an exit report after the evaluation process, which market surveillance authorities and notified bodies will take into account positively, without intervening in their corrective or supervisory powers. Controlled areas are intended to enhance safety, support exchange and cooperation between authorities, the promotion of innovation and competitiveness and learning by testing, and the Commission shall adopt acts specifying the detailed arrangements for the establishment, development, implementation, operation, and supervision of controlled testing areas, with the AIA listing the common principles to be respected, and we should comment on Article 58(f).2, when talking about facilitating the involvement of other actors in the AI ecosystem (Notified Bodies and Standardisation Bodies, SMEs, start-ups, enterprises, innovative players, testing and experimentation facilities, research and experimentation laboratories and EICs, centres of excellence and researchers) to enable and facilitate public and private cooperation.

- The AIA also gives the Committee a number of functions related to advising and assisting the Commission and the Member States in the implementation of the rules in Article 66, giving it the possibility, but not the obligation, to contribute to the coordination and cooperation of market surveillance authorities, collect information, provide advice, issue recommendations, opinions, codes of conduct, evaluation of the AIA itself, and harmonised standards, common criteria, integration of institutions, issuing and receiving opinions, etc. and in particular for the public sector, in paragraph d) *contribute to the harmonisation of administrative practices in the Member States, including in relation to the derogation from the conformity assessment procedures referred to in Article 46, the functioning of AI regulatory sandboxes, and testing in real world conditions* *referred to in Articles 57, 59 and 60;* closing the circle of integrated national and European actors for the implementation of the AIA.

- As we have already pointed out, Art. 79 establishes a system of risk control that may have a particular relevance in cases of uses of AI by public authorities, in relation to the risk notifications in these cases of Art. 82.

- Finally, a brief mention should be made of the administrative sanctions and fines for institutions, bodies, offices and agencies of the European Union in Articles 99 and 100 of the AIA: to ensure the application of the provisions, the text includes a system of sanctions that it is up to the member states to determine, within the margins of this article. The penalties must be effective, proportionate, and dissuasive, taking into account subjective criteria such as the interests of the emerging companies, economic viability, etc. The amounts depend on the seriousness of the infringement, and the possibility of imposing fines in addition to non-monetary measures, such as orders or warnings, is envisaged. With regard to the administrative fine, the specific amount should also be decided depending on the situation, nature, seriousness, delay, etc. Furthermore, it is also foreseen that the European data protection supervisor may impose administrative fines on the institutions, agencies and bodies of the Union falling within the scope of application of the AIA, if not on a par with private bodies, then in the spirit of the sanction that the administration must also bear. They are also graduated taking into account all the relevant circumstances of the situation in question, and in particular according to the seriousness, duration, consequences, number of persons affected and level of damage, degree of responsibility of the body, actions taken to mitigate the damage or the degree of cooperation with the Supervisor and its reporting, as well as similar previous infringements. The amounts of these administrative fines can be up to certain amounts, based on thresholds that depend on the type of non-compliance and are calculated according to the turnover of the company concerned, which will require a differentiated and adapted specification for the public authorities.

III. Some conclusions on the application of the Act to the public sector

The AIA introduces, as we have seen and pointed out from the outset, a flexible system of intervention, which analyses the risk to the fundamental rights and values of the Union as a premise for conditioning the use of AI of the three degrees of affectation, but which is essentially aimed at regulating, ordering and controlling the use of AI, and the possible risks derived from it with respect to the introduction of products or provision of services in the market based on essentially industrial and commercial dynamics. Therefore, it is not specifically oriented towards the regulation, control, and minimisation of risks with respect to the use of these products or services by public authorities, but this does not preclude the application of this regulation in the same terms, and with some of the particularities indicated, when public authorities make use of these systems. Regardless of the likelihood that in the future there will be additional specific rules for administrative or judicial action, due to the additional risks that the use of AI by public authorities entails for the sphere of citizens' rights and duties, this first step, which already introduces important controls, hitherto non-existent, on the use of AI for the adoption, for example, of administrative or judicial decisions (or to assist them), can only be viewed positively.

First, as in the private sector, there are uses where AI is directly prohibited for public authorities. The AIA does not prohibit its use, as some legal systems do, for the adoption of discretionary decisions. However, it does prohibit those uses that could generate a very serious overall impact on rights based on security dynamics that could endorse authoritarian drifts of disproportionate control over the population and which are centred, as can be seen in the AIA, on the functions of public administrations related to security policing, with the evaluation or classification of persons, with "real-time" remote biometric identification systems, as well as with public systems in which their use is only permitted under certain conditions when they do not generate harm or unfavourable treatment or are indispensable for the location of victims, suspected criminals, or the prevention of threats .

One level below the above, we find the high-risk uses, which comprise certain public sector activities where the data subject is usually in an unfavourable and more vulnerable position before the authority (e.g., migration management, asylum and border control, with risk assessment and document verification, or public security, with biometric identification and categorisation of persons, or systems for assessing access to services and benefits). The requirements that this second group must meet will be determined, implemented, and monitored by various entities, with legal figures vital to the functioning of the system such as notifying authorities, the national supervisory authority (which may also exercise the position of market surveillance authority), or controlled testing grounds for AIs, together with the elements of publicity and transparency of registers, and sanctioning systems that ensure compliance with the rule. In addition, and somehow as a certain embodiment of the principle of "reserve of humanity", human surveillance is imposed on AI systems that entail a higher risk to rights (Article 14) and as a result of the extensive list in Annex III, points 5 to 8, we can consider that practically all administrative activities or judicial decision-making that are carried out with the help or entirely by AI and impact on the legal status of citizens, affecting their sphere of rights and duties, will also be considered high-risk. Finally, in addition to specific rules for types of AI such as *chatbots* or equivalent, which the administration will obviously have to take into account when using them, the AIA also makes high-risk systems comparable in practice to general-purpose systems in almost every respect, so that when these are used, in practice, the same precautions must be taken and compliance must be demanded from software providers.

Thirdly, and as is also the case in the private sector, a third group of AI uses are classified as low or non-existent risk (essentially, those that help to improve processes and *back-office*, without direct impact on the status of citizens, to mention the clearest example in the field of public administration), and are considered free to develop and use, without the restrictions of AIA, but without prejudice to the possibility of voluntarily submitting to those foreseen for high-risk systems through codes of conduct.

Finally, a brief reflection on technological innovations and their regulation: the rapid evolution of technology requires a legal framework that is effective but adaptable to constant developments, a feature that is intended to be ensured by the Council's possibility of extending or modifying the content of the AIA in accordance with the vicissitudes that may arise, revisions by way of reissuing the annexes that must be adapted to the changing reality of AI.

In the preceding pages we have tried to systematise the new European Union regulations which, although, as we have already stated at the beginning of the analysis, we believe should not necessarily be regulated in the same regulatory text as that applicable to private entities, at least not in all its effects, it does guarantee us a minimum legal framework to abide by (human control, registers, supervisory bodies, minimum requirements, and above all, risk classification), because as on so many other occasions, society is moving faster than the regulations by which it must be governed. It would be highly advisable in the not too distant future to draw up specific regulations, a specific and appropriate legal regime covering the application of AI for the public sector, which would really help to preserve the specific guarantees that must be safeguarded in this system, in order to have a common European framework in this area, which the public authorities of each Member State would then specify, develop, and adapt to their particularities and domestic law. In this sense, and with regard to the legislation under analysis, what is proposed to be regulated in a harmonised manner, establishing a minimum in terms of control by the Member States, essential guarantees for the fundamental rights, and public freedoms that inform Union law, must be done without imposing such an exhaustive regulation on the Member States that they cannot establish their own rules. And, moreover, it should be emphasised that in no case does the current European regulation in force prevent national legislators from drawing up their own more detailed and guaranteeing internal regulatory bases that harmonise public rights and interests within the framework established by the regulation for the countries of the Union with the protection of citizens' rights. This is, however, a second step for which it was first necessary to make a start. This is what the AIA has done, also for the control of the uses of AI by the public authorities, both administrative and judicial, which already establishes a minimum level of protection that is not negligible and should therefore be highly valued.

LARGE ARTIFICIAL INTELLIGENCE PLATFORMS AND SYSTEMS FOR POLITICAL INFLUENCE: THE INTERSECTION BETWEEN THE "DIGITAL SERVICES ACT" AND THE ARTIFICIAL INTELLIGENCE ACT FROM A RISK PERSPECTIVE

Rosa Cernada Badía

Lecturer in Administrative Law
Catholic University of Valencia San Vicente Mártir

I. Introduction. Basis for the specific treatment of digital platforms and political influence systems in the Act.

It was in the second decade of the 1980s when Ulrich Beck redefined the contours of our modernity by declaring that the class struggle had been overcome and outlining the transition from a society based on the distribution of wealth to a society based on the distribution of risks[1]. In this risk society, conflicts arise from the challenges and conflicting interests derived from scientific-technical development and imply, in the opinion of the German philosopher, a loss of protagonism of states and the birth in their place of "objective communities of threat" that require global solutions[2] .

The idea, which already seemed juicy in the last quarter of the 20th century, is a necessary reference in today's hyper-connected society, in which digital development has reshaped market structures and social levers, provoking an unprecedented legal challenge. In this context, the birth of large digital platforms, as new intermediary bodies, has had a global impact, since: i) it necessarily affects the content and exercise of citizens' fundamental rights[3] ; ii) it has generated new business models, with a transnational dimension[4] under the logic of intermediation and iii) it conditions the relationship between public power and citizens, and may ultimately affect the functioning of democratic systems with phenomena such as disinformation.

[1] Beck, U., *La sociedad del riesgo. Hacia una nueva modernidad*, Ediciones Paidós Ibérica, Barcelona 1998, p. 25.

[2] *Ibid*, pp. 53-54.

[3] European Parliament Resolution of 14 March 2017 on the fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law enforcement (2016/2225(INI)).

[4] Otero Martín, D.; Infante González, J. and Ruiz Mérida, M., "Experiencia comparada: regulación y control de mercados digitales de plataforma en EE UU y China", *Plataformas digitales: regulación y competencia*, n.º 925 (March-April 2022), p. 114.

It is precisely these particular ecosystems of an informational, commercial, and social nature that generate digital platforms on algorithmic structures that are the ideal breeding ground for the proliferation of risks of many different kinds. These risks require a targeted examination of the activity of digital services and the Artificial Intelligence (hereafter AI) systems that underpin them. Beck's idea of threat communities underlies this logic. So does the necessary search for global responses. And the European Union, true to its constitutive principles and its social political vocation, is developing a European response to the challenge of digital governance and the development of AI as a disruptive technology. This response, built on the centrality of the individual as the cornerstone of the digital transition, is part of the so-called European digital strategy[5], which brings together far-reaching regulations. In particular, and together with the General Data Protection Regulation (hereinafter GDPR)[6], the so-called European regulatory package, consisting of the Digital Markets Regulation[7], and the Digital Services Regulation[8] (hereinafter DSA).

The DSA aims to contribute to the proper functioning of the internal market for intermediary services (...) in order to create a safe, predictable and reliable online environment, which promotes innovation and respect for fundamental rights. To this end, it regulates the impact of the state on large platforms as a complement to the strictly private contractual control carried out by them. To this end, based on the *safe harbour*, it establishes a specific liability regime for intermediary service providers. In particular, Title III of the Directive lays down a series of due diligence obligations. These obligations are subject to public supervision through an institutional framework led by the European Commission, which guarantees compliance with the regulations by large technology companies, subjecting them to investigation

[5] European Commission, "Shaping Europes's Digital Future", (February 2020), available at: https://eufordigital.eu/wp-content/uploads/2020/04/communication-shaping-europes-digital-future-feb2020_en_4.pdf, last accessed 15 February 2024.

[6] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of individuals with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation).

[7] Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Regulation) (Text with EEA relevance), OJEU No L 265/1 of 12 October 2022.

[8] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a single market for digital services and amending Directive 2000/31/EC (Digital Services Regulation) (Text with EEA relevance), OJEU L277/1 of 27 October 2022.

procedures and, where appropriate, penalties for failure to comply with their responsibilities.

This regulatory structure has been the subject of attention by the Artificial Intelligence Act[9] , hereinafter AIA, insofar as large platforms base their business model on a fundamental service: personalisation, which makes use of AI systems and models. The risks arising from the incorporation of these systems into the activity of large digital platforms are significant, given the volume of users of these platforms and their potential influence on fundamental rights, online security and the shaping of public opinion.

Alongside this, and closely related to it, the AIA also gives specific treatment to systems aimed at political influence. Its special consideration stems from the European institutions' concern about the development and use of political manipulation techniques using AI systems within the broad framework of the fight against disinformation.

The relationship between disinformation and its effects on politics has been highlighted in various circumstances, although it was particularly glaring in the 2016 presidential election campaign in the United States (*Cambridge Analytica* case) or the BREXIT. The origin of the European reaction stems precisely from an electoral event, the European Parliament elections in May 2019, which prompted this response, particularly with regard to strategic communication and political advertising practices. This highlighted[10] the need for public intervention in this area, prioritising transparency over content procrastination so that users can understand how the output results of their information searches are constructed or how their *feeds* are personalised.

However, systems designed for political influence do not always make use of large platforms to achieve their goals. A recent example is the orchestrated telephone campaign in the United States whereby voters received a call with President Biden's voice urging abstention in New Hampshire[11]. Therefore, alongside the regulation of large platforms, specific attention to this phenomenon is necessary in a growing context of the use of AI techniques to influence election results and, in particular, the targeting of election advertis-

[9] Proposal for a Regulation of the European Parliament and of the Council of 21 April 2021 laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislation, COM (2021) 206 final.

[10] Mardsen, C. and Meyer, T., 'Regulating disinformation with Artificial Intelligence', *Parliamentary Research Service of the European Parliament*, Brussels, European Union (2019), p. 6.

[11] Vid. Doménech, E., "El "deepfake" que imita a Biden en plena campaña alerta a los expertos ante el uso de AI para manipular elecciones", NEWTRAL (21 January 2024), available at: https://www.newtral.es/ia-imita-biden-deepfake-expertos/20240124/, last access, 17 January 2024.

ing, regulated by the European Regulation on Transparency and Targeting of Political Advertising (hereinafter TTPA)[12] .

This paper will therefore examine the specific treatment that the AIA and the TTPA give to large platforms and systems aimed at political influence, assessing their regulation and examining the articulation of the various regulations in force and their systematicity. To this end, we will first examine the *iter legis* of the AIA, analyse the different approaches to the treatment of risk in the applicable regulations and, finally, we will detail the special features provided for in the AIA.

II. A brief look at the "*iter legis*" of the Act on the regulation of large platforms and political influence systems

The Artificial Intelligence Act is part of the European digital market regulation strategy. This was indicated in the Commission's initial text[13], the explanatory memorandum of which referred to the necessary coherence between the future AIA and the Union's services regulation and the DSA (at that time at the proposal stage). In particular, as regards large platforms, the AIA in the common position of the European Council detailed in recital 12 and Article 2.5 the application of the text without prejudice to the provisions relating to the liability of intermediary services[14]. With this brief mention, the relationship between the various pieces of legislation, intended to be applied simultaneously, was being addressed.

However, in the *iter legis* of the proposal and, specifically, in the work of the European Parliament, a special reference is made to large platforms. Indeed, in line with the concern shown in relation to disinformation and the governance of large platforms[15], the explanatory memorandum of the Euro-

[12] Regulation (EU) 2024/900 of the European Parliament and of the Council of 13 March 2024 on transparency and targeting in political advertising, "OJEU" No 900 of 20 March 2024.

[13] Text available at: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF, p. 5, last accessed 12 November 2023.

[14] European Council, 'General Approach on the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Regulation) and amending certain legislative acts of the Union', (6 December 2022), available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=consil%3AST_15698_2022_INIT, last accessed 14 November 2023.

[15] Argelich Comelles, C., "Gobernanza de las plataformas en línea ante la DSA y las pro-

pean Parliament's report[16] includes clarifications of interest for the purposes of this paper. Thus, firstly, the assessment of the AI systems used by candidates or parties to influence votes in elections at all territorial levels as high risk and, together with them, the AI systems used to count these votes. The great potential of these systems is highlighted in parliamentary sessions, as they have the capacity to influence a large number of Union citizens and, ultimately, the functioning of democracy itself. The explanatory memorandum also refers to the relationship between data protection and digital services regulation.

An examination of the parliamentary work in the AIA legislative process[17] does not reveal a direct correlation between the committees' contributions and the final text in relation to the treatment of the major platforms. However, it is worth highlighting the work of the Committee on Culture and Education, whose justification makes explicit reference not so much to large platforms as to their activity. To this end, rapporteur Marcel Kolaja suggests that systems used by the media to create or disseminate automatically generated news articles and AI technologies used to recommend or rate audiovisual content should be considered high-risk systems[18]. While his proposed amendment 55, which suggested algorithmic transparency guarantees on parameters used for content moderation and personalisation, was not accepted, his input was instrumental in the drafting of Annex III, point 8 in the Parliament version.

In particular, from the text approved in Parliament prior to the triduum, two fundamental contributions to the matter in question stand out. These were amendments 739 and 740, which proposed amending point 8 of Annex III by considering systems intended for political influence and recommenda-

puestas de reglamento de mercados digitales e inteligencia artificial (DMA y AIA)", *ADC*, vol. II (April-June 2022), pp. 501-530.

[16] European Parliament, "REPORT on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union" (22 May 2023), *Vid:* Explanatory Memorandum, p. 398.

[17] European Parliament, "Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD) (Ordinary legislative procedure: first reading)", available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_ES.html, last accessed 15 November 2023.

[18] Accessible at: https://www.europarl.europa.eu/doceo/document/A-9-2023-0188_ES-.html#_section4, pp. 452 and 453, last accessed 14 November 2023.

tion systems used by large platforms to be high-risk. The link with the DSA was therefore direct and explicit, but the final reception was mixed.

1. Specific attention in the Act to Artificial Intelligence systems for political influence

Paragraph 8.1(a)(aa) of Annex III of the Parliamentary version of the AIA considered high-risk AI systems those intended to be used for "influencing the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote in elections or referenda". This text has been incorporated unchanged in point 8(b) of Annex III in the final version of the AIA.

In this respect it is essential to follow the wording of the AIA which refers to systems "intended to be used". Here an important nuance should be borne in mind. A system generated with a purpose is not the same as one used for that purpose. This sounds like a mere subtlety, but it is not. Let us look at an example. In the list of prohibited practices in Article 5 of the AIA, the text refers specifically to AI systems that deploy subliminal techniques "with the objective, or the effect of" distorting a person's behaviour. In this case, the text distinguishes these two scenarios in a clear-cut manner. In the case of political influence systems, the wording is less clear as to whether these are systems created for the purpose of political manipulation or whether they include systems which, while not created for political manipulation, can be used for this purpose. Perhaps more clarity could be provided in this regard, as AI systems used for political influence encompass the use of techniques of a diverse nature:

1. micro-targeting techniques[19] or targeting[20] policies that serve as the basis for behavioural political advertising[21] and which would fall squarely under the consideration of high-risk systems in Annex III;

2. the creation of fake profiles on social networks (*bots*) with a certain

[19] Which the European Parliament considers a particularly pernicious form of digital advertising. Mardsen, C. and Meyer, T., *op. cit.*, p. 13.

[20] *Vid.* European Data Protection Committee, Guidelines 8/2020 on targeting users in social media, adopted 13 April 2021, available at: https://edpb.europa.eu/system/files/2021-11/edpb_guidelines_082020_on_the_targeting_of_social_media_users_es_0.pdf, last accessed 20 February 2024.

[21] European Data Protection Board, 'EDPB Urgent Binding Decision on processing of personal data for behavioural advertising by Meta', *Press Room* (1 November 2023), available at: https://edpb.europa.eu/news/news/2023/edpb-urgent-binding-decision-processing-personal-data-behavioural-advertising-meta_en, last accessed 20 February 2024.

ideological profile that generate synthetic information through AI[22]. If these systems are created for political influence, they constitute a high-risk system under the AIA;

3. the automated creation of fake news, which can go as far as ultra-fake news, as was the case in the recent Slovak parliamentary elections[23];

4. techniques not strictly aimed at political influence, but with undeniable impact such as content prioritisation or recommender systems[24].

In view of the final treatment in the AIA of algorithmic content recommendation or prioritisation systems, the ideal interpretation of point 8.b) of the AIA seems to be the strict one, that is, the one referring to systems created with the specific aim of influencing the electoral behaviour of citizens. This interpretation is also in line with the definition of political advertising in Article 3.2.b) of the TTPA, insofar as, in order to be qualified as such, it demands a double requirement: that it may influence the outcome or electoral behaviour "and is designed for that purpose".

In any case, it seems undeniable that the AIA is concerned to link the concept of political influence to the formation of political opinion and the protection of the fundamental rights of political participation of the final recipients of AI systems. This intention is necessarily derived from the final paragraph of point 8.b) of Annex III, which excludes from high-risk consideration systems to whose output information is not directly exposed to natural persons, such as AI systems for logistical and administrative management of political campaigns. For example, systems used to assist in the financing or design of political campaigns (algorithmic campaign advisors)[25] .

Consequently, the treatment of systems aimed at political influence requires consideration of the systematic application of existing rules. In particular, the application of the algorithmic transparency safeguards proposed by the AIA for high-risk schemes in relation to:

[22] Panditharatne, M., "How Artificial Intelligence puts elections at risk and the measures required to protect us", *Brennan Center-Analysis* (21 June 2023, updated 13 July 2023), available at: https://www.brennancenter.org/es/our-work/analysis-opinion/inteligencia-artificial-pone-en-riesgo-elecciones-medidas-proteger-democracia, last accessed 22 February 2024.

[23] Whose impact on Michal Šimečka's defeat of pro-Russian candidate Robert Fico is yet to be determined. Solon, O., "Trolls in Slovakian Election Tap AI Deepfakes to Spread Disinfo", *Bloomberg News* (29 September 2023), available at: https://www.bloomberg.com/news/articles/2023-09-29/trolls-in-slovakian-election-tap-ai-deepfakes-to-spread-disinfo, last accessed 22 February 2024.

[24] For example, systems for personalising political messages, which are based on segmentation and speech adaptation techniques using AI and which reach users through recommendation systems, generating the risk of generating information bubbles.

[25] Scheiner, B., "Six ways AI could change policy", *MIT Technology Revier* (7 August 2023).

i) the GDPR, as data is "the most powerful weapon" to generate political confrontation through profiling and personalisation of information[26];

ii) sectoral regulations; specifically, the DSA's online platform transparency requirements and the TTPA, to which Recital 62 of the AIA refers when declaring the joint application of both regulations. However, it should be noted that the scope of application of the TTPA is not limited to political advertising broadcast on platforms or search engines, but covers all parties involved in the process of preparation, insertion, promotion, publication and dissemination of political advertising (providers or publishers of political advertising and related services).[27]

2. Algorithmic recommendation systems: treatment in the Artificial Intelligence Parliament proposal version and its rationale

The AIA proposal defined algorithmic recommender systems on large platforms as a high-risk system. The basic rationale for this decision rested on the impact of these systems on fundamental rights, a circumstance that the explanatory memorandum itself identified as a "particularly relevant" criterion for their classification as high risk. Furthermore, given the volume of users of large-scale digital platforms, recital 40b of the AIA highlighted their potential influence "on online security, the shaping of public opinion and discourse, electoral and democratic processes and societal concerns" as justification for their classification as a high-risk system.

As noted *above*, the platforms structure their activity around personalisation, which takes the form of the platform's ability to recommend specific content created by the network's users themselves to its users. However, as can be deduced from recital 70 of the DSA, the European legislator expresses a dual concept of recommendation system: (i) the strict one referring to this proposal or suggestion activity and (ii) a broad concept, which would also cover other techniques such as the algorithmic classification and prioritisation of information, the distinction between text and other visual forms or the personalised organisation of information. Precisely in this broad sense, the EDPS recalled that some of these techniques, such as profiling or micro-segmentation, may significantly affect fundamental rights[28].

[26] García Mahamut, R., "Elecciones, protección de datos y transparencia en la publicidad política: la apuesta normativa de la UE y sus efectos en el ordenamiento español", *Revista Española de Transparencia*, n.º 17 extraordinario (2023), p. 78.

[27] *Ibid*, p. 4.

[28] European Data Protection Supervisor, "Opinion 3/2018 EDPS Opinion on online manipulation and personal data" (19 March 2018), p. 9, available at: https://edps.europa.eu/

Despite these considerations, the classification of digital platform recommendation systems as high-risk was not finally accepted in the AIA. In this decision, the legislator has taken into account the relationship between the AIA and the DSA, which is the reference standard for the treatment of digital services. It is therefore appropriate to analyse the intersection between the two rules for the purpose of assessing the treatment of AI platforms and systems used by them in European law and the legislator's final decision regarding their specific treatment in the AIA.

III. The logic of risk on large platforms: the complementarity between the DSA and the AI Act

The current trend in the treatment of large platforms by the European legislator is based on the logic of risk, being this an approach that consists of tailoring rights and obligations to the risks arising from a certain activity[29]. Originally adopted by the GDPR, risk-based compliance inspires the key rules of European digital law, in particular the Regulatory Package and the AIA, albeit from a different perspective.

Indeed, without disregarding the undeniable benefits that large platforms bring to citizens, the fact is that their misuse can have a profound impact on fundamental rights and democratic systems that should be addressed[30], given the digital mutation of the information markets mentioned *above*. The treatment of risk in the DSA is also based, albeit indirectly, on a certain scale of risks that is articulated on the basis of a cumulative requirement of obligations. To this end, it distinguishes between five scales of obligations in Chapter III:

1. the general provisions applicable to all providers of intermediary services (Section 1 Articles 11 to 15);

sites/edp/files/publication/18-03-19_online_manipulation_en.pdf, last accessed 21 November 2023.

[29] Barrio Andrés, M., "El cumplimiento basado en el riesgo o *risk based compliance*, pieza cardinal del nuevo Derecho digital europeo", *Análisis del Real Instituto Elcano (ARI)*, n.º 34 (2023), p. 3.

[30] It is a matter of going beyond a strictly technological examination to qualitatively assess the impact of the activity of these platforms on the rights of individuals (*Risk to rights approach*, taken from data protection law), *vid.* MAHLER, T., "Between risk management and proportionality: The risk-based approach in the EU's Artificial Intelligence Act Proposal", *Nordic Yearbook of Law and Informatics* (September 2021), available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4001444, p. 259, last accessed 17 January 2023.

2. additional provisions applicable to providers of hosting platforms, including online platforms (Section 2, Articles 16 to 18);

3. the obligations applicable to providers of online platforms (section 3, articles 19 to 28);

4. as a consumer protection speciality, the additional provisions in Section Four (Articles 29 to 32) concerning online platforms enabling B2C business (*Business to Consumer)*;

5. obligations for providers of very large online platforms (section 5, articles 33 to 43).

The regulatory perspective of the DSA is asymmetric, requiring compliance with additional reporting, transparency and *accountability* obligations by very large online search engines and platforms (hereinafter VLOPs). VLOPs are defined in Article 33 as those with more than 45 million monthly active users in the EU (or 10% of the EU population). Under the mandate of Article 33.4, in April 2023 the European Commission adopted the decision designating large platforms, with no little controversy[31].

Recital 75 of the DSA justifies this decision on the basis of the number of recipients of the service and the central position of these systems in facilitating the exercise of freedom of expression and information and the shaping of public opinion. The DSA takes care to base the proportionality of these measures precisely on the assessment of the risks arising from VLOPs (recital 76) by correlating the intensity of the measures with the social impact of this type of platform, which ultimately has the necessary resources to carry out an assessment of the risks they give rise to (*ex* Article 34) and to address their consequences[32].

These risks have come to be defined as systemic risks, which arise from the design or operation of the service "and related systems, including algorithmic systems". Systemic risk refers to a holistic perspective insofar as risks to human health, the environment or fundamental rights are embedded in a broader context of social, financial, and economic risks and opportunities,

[31] The designated platforms have expressed their disagreement. The list is available at: https://ec.europa.eu/commission/presscorner/detail/es/ip_23_2413 Rumours have even gone viral about the reluctance of some networks to comply with the DSA's requirements, although in reference to X (former Twitter) they have been denied: *Vid.* https://www.lavanguardia.com/tecnologia/20231023/9320565/elon-musk-desmiente-rumores-eliminar-twitter-paises-ue.html, last access, 12 November 2023.

[32] Castelló Pastor, J. J., "Nuevo régimen de responsabilidad de los servicios digitales que actúan como intermediarios a la luz de la propuesta de Reglamento relativo a un mercado único de servicios digitales", in Castelló Pastor, J. J. (dir.), *Desafíos jurídicos ante la integración digital: aspectos europeos e internacionales,* Aranzadi-Thomson Reuters, Cizur Menor (Navarra) 2022, p. 73.

combining natural phenomena, socio-economic developments, technology and multi-level policy actions[33]. Article 1.44. of the AIA also defines systemic risk in this sense by reference to the significant impact that general purpose models (hereinafter GPM) may have on the internal market, in particular "actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights or society as a whole, which may spread at scale through the value chain".

Therefore, an analysis of systemic risks requires a three-pronged process of risk identification, assessment and management from a multidisciplinary perspective that allows for an analysis of the interdependencies and relationships between various risk groups. Article 34 of the DSA requires VLOPs to assess the following systemic risks:

(i) the dissemination of illegal content through their services, e.g., relating to hate speech and disinformation;

(ii) any actual or foreseeable negative effects for the exercise of fundamental rights. The text expressly refers to the risks arising from the design of algorithmic systems of VLOPs aimed at limiting freedom of expression (automated content moderation);

iii) any actual or foreseeable negative effects on civic discourse and electoral processes, and public security;

(iv) any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors and serious negative consequences to the person's physical and mental well-being.

This list does not constitute a *numerus clausus*, insofar as it concretises the general clause of the first indent requiring the detection, analysis and assessment of "any systemic risk" in the Union arising from the design or operation of its service and the systems used, with particular reference to algorithmic systems.

It can be seen that the systemic risks identified in the DSA are based on a guarantee stance, on strengthening the position of the natural person receiving the service. It is worth noting at this point the different approach of the DSA and the AIA. While the DSA is oriented towards guarantees for platform users who, in the final analysis, are the final recipients of the AI systems used by them, the AIA is based on the role of providers, deployers, importers, and distributors of systems who do not necessarily have to be their final re-

[33] Renn, O. and Klinke, A., "Systemic risks: a new challenge for risk management", *EMBO Reports*, Volume 5, Special Issue (October 2004), p. 41, available at: https://doi.org/10.1038/sj.embor.7400227, last accessed 27 February 2024.

cipients[34]. These approaches are complementary in the overall assessment of the risks underlying the use of AI systems on digital platforms.

Indeed, the general risk perspective that inspires the DSA differs from the AIA scheme. The AI regulation does not leave risk assessment to the obliged parties, but imposes a legislative analysis of risk accompanied by management systems by the obliged parties. This *risk-based approach* in the logic of the AIA has therefore come to be seen more as a legislative technique for limiting the scope of application of the regulation and ensuring legislative proportionality[35]. In any case, the solution proposed in the parliamentary version involved incorporating the protection of the natural person user into the logic of a regulation centred on the platform as the deployer/developer. The need to bring this perspective into the standard seems absolutely appropriate, insofar as it responds to the primary objective of Article 1 of the AIA. The user of VLOPs in his own right, and also on axiological grounds, deserves to be taken into account in the regulation of AI systems. The nuance is of a systematic nature and not only in relation to the AIA itself but, beyond that, in the desirable systematisation of European digital law. Therefore, in plain English, the question is not whether users should be taken into account, but where and how.

The fact that in the AIA the qualification of algorithmic recommendation systems for VLOPs as high risk has been dropped could be seen *a priori* as a loss of person-centredness, a sort of betrayal of the basic objective of the regulation. However, a systematic examination of the regulation in relation to the requirements of the DSA indicates that this treatment is ultimately more congruent with the European system of digital law in considering the intensity of the risks generated by VLOPs.

Indeed, one cannot fail to recognise that the European legislator, since the Commission's version of the AIA, had the end recipient of AI systems in mind in the original Article 52 when regulating transparency obligations for providers or implementers of AI systems "intended to interact directly with natural persons". The obligations, focused on algorithmic transparency, sought to ensure that the end recipient was informed that it was interacting with an AI system.

In addition to these provisions, the final text has refined the concept of GPM, incorporating a sort of intermediate system between GPM and high

[34] Jiménez-Castellanos Ballesteros, I., "Decisiones automatizadas y transparencia administrativa: nuevos retos para los derechos fundamentales", *Revista Española de la Transparencia*, n.º 16 (2023), pp. 202-203.

[35] Mahler, T., *op. cit.*, p. 247.

risk. These are the models of general use with systemic risk (GPMSR), provided for in article 51, a new version of the text. These models are declared as such by the European Commission:

- on the basis of their high impact capacities, to be assessed using appropriate tools or methodologies, in accordance with the criteria in Annex XIII.

- or depending on their technical computing capacity or processing power. Article 51.2 specifies this criterion by defining as systemic risk those models whose cumulative amount of computation used for training measured in floating point operations (FLOPs) exceeds 10^25. This would include, for example, Chat GPT-4.

- However, paragraph 3 allows the Commission to adjust these thresholds by means of acts of conformity in order to take account of technological developments.

If qualified as GPMSR, AI models must meet some of the obligations of high-risk systems, in particular: (i) conduct a model assessment in accordance with standardised protocols and tools; (ii) assess and mitigate EU-wide systemic risks; (iii) conduct documented follow-up of serious incidents and corrective actions, communicating these circumstances without undue delay to the AI Office; and (iv) ensure an adequate level of cybersecurity protection for the model.

The fact is that the provision is sufficiently flexible to adapt the rule to technological developments. But beyond embracing the technical contingency, the key to analysing its applicability to VLOPs lies in understanding the impact of this new category in the AIA risk system. It should be borne in mind that the standard speaks of high-impact capacity, not of the impact actually produced. In other words, it is strictly assessing risks and subjecting a particular type of risk to specific obligations. The qualification of the original risk grading system, which in the final text is disguised as a specification in the GPM classification, directly affects large platforms. Thus, the joint application of both regulations seems to lead to the question of whether a VLOPs platform can be incorporated into the GPMSR category by the European Commission. However, a careful examination of this question invites a reformulation of the approach.

In this regard, it is worth mentioning recital 118 of the AIA which, in view of the obligations imposed by the DSA, considers that the obligations of the AIA must be deemed to be fulfilled unless systemic risks not covered by the DSA are identified. Therefore, the issue here is not so much to follow the self-evident logic of Article 51, i.e., the possible classification of AI models that may incorporate VLOPs as GPMSR, but to assess the risks actually covered by the platforms in compliance with the DSA's risk management model in relation to the AIA. This issue will be returned to in the next section when examining the specialities of AIA implementation in terms of risk management.

IV. Special features of the application of the Act to large-scale AI platforms and systems for political influence

As regards the application of the AIA in the digital platform sector, we must start from Recital 118 of the AIA which recognises the complementarity with the DSA as regards the obligations imposed on intermediary service providers. Therefore, based on the joint application of the two pieces of legislation, the key question is to dissect the safeguards that the DSA imposes on the AI systems used by platforms and to examine their impact on the AIA. The transparency requirements of the TTPA with respect to systems for political influence will also be taken into account.

The mandatory package that the DSA imposes in its relationship with the AIA is based on a general duty of compliance taking into account the following aspects: the generally recognised state of the art, the purpose of the system, foreseeable misuses and the system of risks. On the basis of this general scheme, its joint application with the AIA is conditioned by the subjective scope of application of both rules. Thus, according to the terminology of Article 2.1 of the AIA, platforms can be providers[36] or responsible for the deployment[37] of the AI systems they contract to provide their service. On this basis, the joint application of the AIA, the DSA, and the TTPA can be examined in the light of the safeguards that the DSA imposes on AI systems used by VLOPs and taking into account the specialities arising from the use of AI systems for political influence by platforms. These safeguards can be classified into four blocks: i) risk management safeguards; ii) algorithmic transparency safeguards; iii) procedural safeguards; and iv) organic safeguards.

1. Risk management safeguards

With regard to the risk management system, as noted above, the AIA subjects political influence systems to the risk management procedure in Ar-

[36] This is the case of Meta and its platforms, *see* https://ai.meta.com/blog/powered-by-ai-instagrams-explore-recommender-system/, last accessed 28 February 2024.

[37] The AIA Council version extends the obligatory scope of providers to the users of AI systems, a figure that in the Parliament version is redefined as an implementer. The final text qualifies the subjective scope by referring to providers and those responsible for deployment, without prejudice to its application to other subjects such as importers or distributors.

ticle 9 of the AIA as high-risk systems. The TTPA also takes into account the wide range of political advertising services that can generate risks. Thus, to the extent that VLOPs provide their services as publishers of such advertising, they are subject to the DSA's risk management system by express reference in recital 46 of the TTPA. The text does not clearly delineate the cases, but a systematic reading of the three rules allows us to understand this reference to the DSA as referring to political advertising services that are not considered high-risk by the AIA (for example, if they do not use AI systems).

For its part, and with respect to AI systems or models embedded in VLOPs, the AIA refers to the risk management framework of the DSA and, to this end, presumes that the obligations of the AIA are met unless significant systemic risks not covered by the DSA arise or are identified. This reference constitutes a *rebuttable* presumption of compliance with the standard imposed by the AIA, subjecting insufficient coverage to the heightened obligations of Article 55. Therefore, what is significant is not the qualification of the AI models incorporated in the VLOPs as GPMSR but the adequate coverage of risks in the application of the DSA and the AIA management system.

In this regard, recital 118 of the AIA requires VLOPs to assess the potential systemic risks arising from the design, operation and use of their services and, in particular, extends this assessment to (i) algorithmic systems used in the service that may contribute to these risks and (ii) systemic risks arising from possible misuses. The connection with Section 5 of the DSA is clear and, in particular, with Article 34, the content of which it reproduces. This risk assessment duty arises at a very specific point in time: when designated as VLOP and in any case once a year or before deploying functionalities that may have a critical impact on risks.

The risk assessment should take into account the purpose of the system. In this respect, the DSA qualifies the factors to be assessed, which include: (i) the design of recommendation systems; (ii) content moderation systems; (iii) applicable general terms and conditions and their enforcement; (iv) ad selection and display systems; (v) effects of political advertising services, by reference to the TTPA, already cited; and (vi) data-related practices of the provider.

Along with these issues, and similar to the parliamentary version, the DSA calls for an investigation into some abuses, namely those that happen because of manipulating the service, such as automated exploitation or fake use (bots). Furthermore, the harmful effects of appropriate uses of AI must also be assessed where the wrongfulness lies in other aspects of the service, such as the potentially rapid and wide amplification and dissemination of illegal content (viralisation) or information incompatible with the general terms and conditions. This same logic applies to targeted advertising practices and other techniques limited by Article 18 of the TTPA.

As has been argued, the logic of the AIA is preventive, therefore, the risk assessment requires the correlative adoption of "appropriate and specific" measures to minimise risk and facilitate adequate and proportionate compliance with the requirements of Chapter II. The same logic is adopted by Article 35 of the DSA, which requires VLOPs to adopt reasonable, proportionate, and effective risk mitigation measures, tailored to systemic risks and taking into account the consequences of such measures on fundamental rights. These measures are monitored by the European Commission with the invaluable technical support of the European Centre for Algorithmic Transparency (ECAT).

Therefore, in both texts there is a systematic vision of the corrective measures, which operate in two moments: i) from the design and development of the system and ii) once designed, as mechanisms for control and mitigation of non-eliminable risks (think of the dissemination of electoral advertising on days of reflection), all of this accompanied by the necessary algorithmic transparency and literacy of those responsible for the deployment of the system.

This systematic view of remediation mechanisms invites, in the context of the DSA, to take into account the application of specific measures. For example, the testing of algorithmic systems in Article 35 of the DSA[38], the scrutiny of which Article 40 specifies by requiring platforms to explain to the Commission or the DSA coordinator the design, logic, operation, and testing of their algorithmic systems. For its part, once the system is in place, the adaptation of algorithmic systems, including recommender systems.

Article 34 of the DSA requires large platforms to keep supporting documents of risk assessments for at least three years. This is substantially less than the 10-year period foreseen for high-risk systems in Article 18 of the AIA. Given that the AIA does not provide for any minimum retention period for GPMs, this intermediate period, which reflects the systemic view of risk, seems reasonable in terms of the regulatory system.

2. Algorithmic transparency safeguards: explainability vs. opacity

With regard to algorithmic transparency obligations, the different focus in subjective terms of the AIA and the sectoral regulation we are analys-

[38] Requirement that the AIA relies on high risk systems ex Article 9.5 to 9.7.

ing should be taken into account. Therefore, their examination must be approached from a broad perspective, in terms of the value chain, with the AIA focusing on the providers/deployers and the DSA and TTPA on the final recipients. Herein lies the proper articulation of transparency as the ultimate guarantee of the explainability of the AI systems used by VLOPs in European digital law.

Indeed, the DSA and the TTPA focus on the end-user of the platform or recipient of the advertising, however, this perspective is also addressed by Article 50 of the AIA which sets out transparency obligations for providers and users of AI systems intended to interact with natural persons. The algorithmic transparency guarantees of the DSA stem from Article 14 which provides for a kind of contractual algorithmic transparency by requiring platforms to include information on automated decision-making in their general or service provision terms and conditions. This requirement can be reflected in Article 50.2 of the AIA by requiring that, at the latest, information on the use of algorithms in networks is provided to natural persons at the time of the first interaction or exposure. Explainability of the system is ultimately ensured by Article 50.5 by: (i) the accessibility of the information under the last indent of the provision, supplemented by the reference to child-friendly explainability in Article 14.3 of the DSA; (ii) appropriate drafting, which the AIA defines in terms of "clear and distinguishable" wording[39]; (iii) the literacy measures in Article 4 of the AIA.

A specification of this guarantee of algorithmic transparency is contained in Article 27 of the DSA, which requires VLOPs to include in the general terms and conditions information regarding the main parameters used in their recommendation systems and the options that the platform makes available to the recipients of the service to modify or influence these parameters. This information relates to the explainability of the system as it implies the motivation of the decision, i.e., the explanation of why a certain content is suggested. Article 27.2 guides platforms in fulfilling this obligation by identifying two minimum parameters: (i) the criteria which are most significant in determining the information suggested to the recipient of the service, and (ii) the reasons for the relative importance of those parameters.

Further to the protection of the final recipient, Article 52.3 of the AIA regulates the specific case of ultra forgeries, known as *deep fakes*, which may be used, where appropriate, as elements of political influence. These techniques, insofar as they are not strictly for the purposes set out in point 8(b)

[39] Article 14.6 of the DSA requires the publication of the general conditions in all official languages of all Member States in which VLOPs provide services.

of Annex III, are also considered *a priori* to be of limited risk, although the AIA subjects them to additional transparency obligations. In particular, for VLOPs, the technique of labelling *(flagging)* so that users are aware that they are counterfeit. Furthermore, the second subparagraph refers to fake news (text manipulated in order to inform the public on matters of public interest). In this case, labelling is the key guarantee of transparency in the protection of end-users of VLOPs services.

In the specific context of advertising, the AIA labelling requirement is in addition to the requirement strictly identifying the political nature of the advertisement contained in Article 11 of the TTPA and Article 26.1(a) of the DSA. Article 26.1(d) of the DSA requires VLOPs to provide, in a simple manner, meaningful, accessible and direct information from the advertisements themselves about the main parameters used to determine the target of the advertisement and, where appropriate, how to change those parameters. However, as far as we are concerned here, Article 26.3 prohibits online platforms from presenting targeted political advertising, i.e., based on profiling[40] on the basis of personal data revealing political opinions or the other special categories of data in Article 9 of the GDPR. Here, the prior transparency guarantees in Articles 6 and following of the TTPA must be taken into account, in particular Articles 7, 11, 12 and 19 and the requirements for the targeting of political advertising in Article 18 TTPA as well as the duty of retention of political advertisements and transparency notices and their amendments, which the TTPA extends to 7 years from the delivery or dissemination of the advertisement (Article 9.3) or from the last publication of the notice (Article 12.4).

Article 15.1 of the DSA enshrines the information transparency obligations of intermediary service providers in terms of *accountability*, and makes a threefold reference to the transparency of AI systems by requiring accountability through the issuance of reports including, inter alia, information regarding:

(i) the number of notifications processed by automated means only and the average time required to take action (paragraph (b));

(ii) automated content moderation and the type of measures taken affecting the availability, visibility, and accessibility of information provided by recipients of the service and other related restrictions (paragraph c). This

[40] Article 4.4 of the GDPR defines profiling as 'any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person', in particular, for this case, to analyse or predict how that person will vote.

includes recommender systems using AI systems, which must be identified by the provision;

iii) the use of automated means for content moderation purposes (paragraph e). The provision itself specifies the minimum information to be published: qualitative description, specification of the precise purposes, indicators of the accuracy and possible rate of error of the automated means used to fulfil those purposes, and any safeguards applied. With regard to VLOPs, Article 42.2 requires these indicators of accuracy and related information to be broken down by each official language of the Member States. Together with Article 42, the provisions of Article 15 are complemented for VLOPs by Article 24.1, which, as far as we are concerned here, refers, among other issues, to the publication of the number of account suspensions, for example, *bots* used as a mechanism for political influence. In this area, *ex post* transparency is completed by Article 14 of the TTPA with regard to information on the use of targeting techniques.

With regard to the timing of the reports, the annual frequency generally recognised in Article 15.1 of the DSA is qualified in Article 42.1 of the DSA by requiring their publication six months after the platform has been notified of its designation as a VLOP and, once this has been done, at least every six months. The duty of accountability for VLOPs takes the form of a duty to submit to the Commission and the DSA coordinator the Article 42.4 reports, in particular the results of the risk assessment and the specific mitigation measures, the audit report and the audit implementation report and, where appropriate, the report on the consultations the provider has carried out in support of the risk assessments and the design of mitigation measures. These reports shall be publicly available unless there is a reasoned request for the platform to be made fully accessible. This obligation therefore reinforces the guarantees for the user of the system and the final recipient of VLOPs without the need to consider them as high-risk systems.

A specific manifestation of *ex post* algorithmic transparency in the field of advertising lies in the duty to publish and update a repository containing basic and easily accessible information about advertisements advertised on platforms, in accordance with the content of Article 39.2 of the DSA and 13 of the TTPA. In particular, the personalisation parameters of the advertisements (which may make use of AI systems), i.e., the criteria used for the presentation or exclusion of the advertisement to certain users, must be made public. In any case, this duty of information starts with the presentation of the ad on the platform and continues until one year after the last time it was presented.

These algorithmic transparency measures entail the right of the recipients of the service, i.e., the users of the platforms, to know that they are interacting with an AI system. This right is most emphatically recognised in Article 50.1 of the AIA, which requires this guarantee of transparency from the design or development of the system. It is remarkable the nuance that the parliamentary version incorporated by resting this duty of information not only on the system but also on the provider or user. In this way, the platforms (whether they were considered to be providers, users or implementers of the system) were called upon to fulfil this duty of communication[41] to the natural persons using their services. This nuance is dropped in the final text of the AIA, which refers only to system providers and exempts the duty of information where it is evident due to the circumstances "from the point of view of a reasonably well-informed, observant, and circumspect natural person". This wording does not effectively reduce the transparency guarantees for users of VLOPs insofar as the duty of contractual transparency in Article 14 of the DSA and the algorithmic transparency obligations in Article 27 of the DSA subject the platform not acting as a provider to this duty of information.

3. Procedural safeguards

In line with international law, the DSA is based on the position that platforms are exempt from liability for unlawful content uploaded by their users unless they have actual knowledge of the infringement (*safe harbour*), although it recognises the good Samaritan clause. Consequently, the regulation itself allows for the possibility for platforms to take various measures against misuse, including: (i) blocking, relegation of information, (ii) suspension or termination of service for certain users, (iii) suspension or termination of accounts, or (iv) suspension, termination or restriction of monetisation of accounts, in accordance with Articles 3(t) and (17). These decisions may be taken in an automated manner through the use of AI systems which shall comply with the requirements of the GPMs of the AIA. In addition, their activity shall be subject to independent external audits at least annually in accordance with Article 37 of the DSA.

As has been noted, the DSA strengthens the position of platform users and provides them with specific means of intervention and management that may be relevant for the purposes of the AIA. Firstly, Article 16 of the DSA establishes a system of notifications that it requires hosting service providers to set up. Through this system, any user (natural or legal person) or

[41] European Commission, Directorate-General for Communication Networks, Content and Technologies, 'Ethical guidelines for trusted AI', *Publications Office* (2019), p. 22, available at: https://data.europa.eu/doi/10.2759/14078, last accessed 3 March 2024.

in particular a reliable Article 22 alerter can report the detection of illegal content. More specifically, when managing this system of notifications, Article 20 provides, in respect of online platforms, for the establishment of an internal complaint management system which can be automated. Through this system, platforms set up a procedure for resolving complaints against decisions to moderate content contrary to the general terms and conditions of service or to detect political advertisements which breach the provisions of the TTPA.

The processing of these notifications must be carried out "in a non-discriminatory, diligent and non-arbitrary manner" in accordance with Article 20.4, complying with the requirements of the GPMs of the AIA. Therefore, when assessing the output result of an AI system for the automated handling of these complaints, adequate explainability allows the user to understand the parameters followed by the platform in its decision and, where appropriate, to proceed as is in his best interest. In this respect, the reference to the guarantee of algorithmic transparency is key.

This internal system for managing notifications implies placing the platforms in a quasi-judicial position, nuancing the return to public law in the governance model. However, this system is without prejudice to the platform user's possibility of going to an out-of-court dispute resolution system or the corresponding ordinary courts. A notification system similar to that of the DSA, with the possibility of automated processing, is provided for in Article 16 of the TTPA for the identification and, where appropriate, withdrawal of notices that do not comply with the requirements of the TTPA.

In any case, the work of homogenisation in this area carried out by the European Parliament in the *iter legis* of the AIA proposal cannot be overlooked, by extending the right to information to this complaints management system in the wording it proposed for Article 52.1 and (3b), in relation to Articles 27 and 38[42] of the DSA. These provisions are dropped in the final text, although recital 170 recalls the existence of effective remedies for natural or legal persons under European law when their rights or interests are affected by an AI system and recalls the possibility of lodging a complaint with the market surveillance authority in the event of infringement of the provisions of the AIA.

A final note of interest should be made regarding the complaints management system. In the processing of this guarantee, a lack of conformity of the AI system used by the platform may be revealed, in which case, in

[42] This refers to the obligation for large platforms to enable at least one option for recommendation systems not to be based on profiling.

the management of these complaints, the collaboration mechanism for the adoption of corrective actions can be articulated, particularly interesting if the complaint is made by the alerters of article 22 of the DSA.

Finally, the DSA is based on the need to combine the use of algorithmic systems in the provision of services by platforms with human review. This issue is not specifically regulated in the final text of the AIA, which makes no provision for GPMs[43]. However, the guarantee enshrined in the DSA is appropriate insofar as it refers to the need for human supervision of two types of decisions:

- automated moderation of content, in accordance with Article 14 of the DSA. This provision should be seen in conjunction with Article 42.2, which complements the provisions of Articles 15 and 24.1 by requiring *ex post* information transparency (*accountability*) of the data relating to the human resources assigned to this review task when moderation is automated.
- the supervision by qualified human personnel of decisions taken under the complaints system, in accordance with Article 20.6.

The DSA refers to this involvement of natural persons in the supervision of the functioning of algorithmic systems without elaborating on its optional or mandatory nature, although recital 58 seems to imply a certain mandatory nature insofar as it obliges platforms to establish internal complaint handling systems "which are subject to human review where automated means are used".

4. Organic Guarantees and Digital Governance

European digital law has moved from a model of self-regulation to a model of co-regulation, in which elements of *soft law* (codes of conduct or good practices[44]) are combined with an institutional structure that materialises the return to public law, to supervise compliance with the regulations and consolidate the model of digital governance in the Union. To this end, it articulates a series of organic guarantees that take the form of the appointment of authorities and other subjects with supervisory powers. This scheme is followed in the three regulations we are analysing. The relationship between the DSA and the AIA with regard to these organic safeguards was made clear in Recital 40b of the Parliament's version of the AIA by stating, from the

[43] With regard to systems for political influence considered to be high-risk, Article 14 provides for this.

[44] Articles 45 to 47 of the DSA and Article 56 of the AIA, by reference in the scope of this study to Articles 53.5 and 55.2.

necessary impact perspective, that the authorities designated under the DSA were to act as enforcement authorities for the purposes of compliance with the AIA. This provision, however, can be understood as subsumed under the aforementioned complementarity clause in the AIA.

According to the provisions of Chapter IV of the DSA (Articles 49 and following), the competent authorities for the supervision of supervisory service providers and the implementation of the DSA are the digital services coordinator and the European Commission, which exercises important supervisory tasks and may adopt implementing acts. The interlocking of these figures takes place as follows:

- The European Commission is called upon to participate in the AI Governance structure, through the European AI Office, which will collaborate with the European AI Council and which, in accordance with Article 68, may assume the exclusive competence of the Commission to oversee the fulfilment of the obligations of the systems and GPMs.

- For its part, the National Commission for Markets and Competition has been designated as the coordinator for digital services in Spain,[45] which should, where appropriate, coordinate with the Spanish Agency for the Supervision of Artificial Intelligence as the national authority that monitors compliance with and enforcement of the AIA.

- In any case, and given the centrality of risk management and algorithmic transparency in both regulations, the importance of the ECAT (already mentioned), a body based in Seville and specialised in the multidisciplinary analysis (technical, scientific and legal) of the use of algorithms, their risks and impact, should be highlighted. ECAT is a key support for the Commission in examining the transparency and risk self-assessment reports of VLOPs as well as in the practice of implementing acts, especially investigative measures[46]. It will also collaborate, inter alia, with the scientific panel of independent experts foreseen in the AIA.

This institutional structure should be understood without prejudice to other figures, provided for in the DSA for coordination purposes (European

[45] Comisión Nacional De Los Mercados Y La Competencia, "El Ministerio para la Transformación Digital y de la Función Pública designa a la CNMC como Coordinador de Servicios Digitales de España", Press release (24 January 2024), available at: https://www.cnmc.es/sites/default/files/editor_contenidos/Notas%20de%20prensa/2024/NdP-CNMC-DSA.pdf, last accessed 6 March 2024.

[46] European Commission, 'Implementing the Digital Services Act: Commission launches European Centre for Algorithmic Transparency', *Press Release* (17 April 2023), *see:* https://ec.europa.eu/commission/presscorner/api/files/document/print/es/ip_23_2186/IP_23_2186_ES.pdf, last accessed 7 March 2024.

Digital Rights Board); or as a hinge between users, authorities and platforms (appointment of contact points between platforms and authorities in Article 11) and between platforms and end recipients of the service (Article 12); of legal representatives of the platforms (Article 13 of the DSA or 14 of the TTPA)[47] or of heads of compliance verification (Article 41) that ensure compliance with the DSA. This structure relates to that set out in Article 15 of the TTPA which, without prejudice to the appointment of competent authorities for areas not regulated by the DSA, subjects the supervision of compliance with the TTPA to the institutional set-up of the DSA in respect of intermediary services.

V. Conclusions

1. The AIA's treatment of VLOPs and political influence systems is appropriate in systematic terms and responds more precisely to the logic of the applicable regulation, focusing on the provider/deployer in the case of the AIA and on the end-user in the case of the DSA and the TTPA. These different perspectives are complementary in the overall assessment of the risks underlying the use of AI systems on digital platforms.

2. The final treatment of VLOPs, as they are no longer considered a high-risk system in the AIA, is more congruent with the European system of digital law in the consideration of the intensity of the risks generated, insofar as the DSA and the TTPA enshrine a strongly guaranteeing regulation based on transparency.

3. With respect to political influence schemes, the consideration of high risk should be interpreted as limited to those schemes developed specifically for this purpose. Therefore, systems that serve this purpose in an accessory manner (such as recommendation systems or systems for the creation of ultra forgeries) are subject to the regime of Chapters IV and V of the AIA.

4. Article 50 of the AIA welcomes the end-user orientation of systems intended to interact directly with natural persons (such as VLOPs), with Article 51 and following providing for additional obligations if systemic risk is present (GPMSR). The recognition of GPMSR incorporates in practice a new gradation of risks that articulates, thanks to its specific regime of obligations, an intermediate intervention regime between general purpose systems and models and high-risk systems. This intermediate regime is aligned with the mandatory regime of the DSA and the transparency regime of the TTPA.

[47] Corresponding to the authorised representatives of suppliers under Article 54 of the AIA.

5. The recognition of GPMSR status has a tangential impact on the VLOPs regime. The AIA does not require GPMSR categorisation of AI models incorporating VLOPs, but rather an assessment of compliance with the AIA's risk management standard, which is presumed *rebuttable* for VLOPs subject to the DSA's risk management system. The presumption operates except for those systemic risks that are shown not to be covered, in which case the AIA's cumulative obligations would be enforceable. In the same vein, the DSA's risk assessment and mitigation, monitoring and corrective action obligations are comparable to those of the AIA's GPMSR in terms of collateral.

6. The key safeguards rely on algorithmic transparency and explainability of the system provided for in sector-specific regulation, which are complemented on technical issues by the specifications of the AIA, in particular: (i) the information in Annexes XI to XIII; (ii) the reporting duty to be interacting with an AI system (Article 50.1); (iii) the labelling of content generated by AI systems; and (iv) the accountability regime.

7. Beyond the procedural guarantees of the DSA to strengthen the position of the end-user, in the field under study the AIA governance system maintains the leading role of the European Commission (supported by the ECAT) and responds adequately to the co-regulatory regime with the incorporation of national and European authorities, platforms and reliable alerters with a renewed role for public law intervention techniques in a technological and humanistic perspective.

General regime applicable to high-risk Artificial Intelligence systems

THE IMPLEMENTATION OF HARMONISED STANDARDS AND COMMON SPECIFICATIONS IN THE FIELD OF ARTIFICIAL INTELLIGENCE (ARTICLES 40 AND 41 AIA)

Vicente Álvarez García
Professor of Administrative Law at the University of Extremadura

I. Introduction

1. The regulation of Artificial Intelligence through the harmonising technique of the new approach

The regulation of Artificial Intelligence in the European Union has followed the technique of the new harmonising approach, in such a way that the European Institutions have renounced the regulation of this field on their own, in order to appeal to the collaboration of private subjects[1].

This system of public-private collaboration means that the Community public authorities will establish the general regulatory framework to which this product will be subject, including the essential requirements that it will have to meet in order to be validly introduced and marketed on the European market, but the technical specifications that serve to meet these requirements will be established, as a general rule, by the European standardisation bodies. It will also be private parties that will carry out the conformity assessments (or, if you like, the controls) that will make it possible to verify whether the product has been developed in accordance with the relevant technical specifications and, ultimately, whether it complies with the mandatory essential requirements in terms of health, safety, fundamental rights and European values imposed by the legislative act regulating Artificial Intelligence.

The decision to follow the new approach regulatory model means that, in order to address the regulation of Artificial Intelligence, consideration must first be given to its regulatory legislative act, which takes the legal form of a Regulation, with the legal basis provided by Article 114 TFEU, given that

[1] On the application of the new harmonising approach technique in the field of Artificial Intelligence, see, for example, the following two studies: Álvarez García V. and Tahiri Moreno, J., "La regulación de la inteligencia artificial en Europa a través de la técnica armonizadora del nuevo enfoque", *Revista General de Derecho Administrativo*, n.º 63, 2023; and Álvarez García, V., "Los instrumentos normativos reguladores de las especificaciones técnicas en la Unión Europea: un breve ensayo de identificación de nuevas fuentes del Derecho", *Revista General de Derecho Administrativo*, no. 64, 2023.

the purpose of this rule is to "adopt measures to ensure the establishment and functioning of the internal market". It is true that, historically, the form followed by the new regulatory acts regulating products was the Directive, but nowadays the Act has become the norm, due to the need to achieve regulatory uniformity of the goods regime within the Union.

The AI Act is not, however, the only legislation regulating this product, but, secondly, a number of cross-cutting pieces of legislation of the first order apply, which constitute what is known in EU jargon as the "new legislative framework for the marketing of products". Of all this legislation, the most important for this chapter is the European Standardisation Regulation of 2012[2], although in the field of technical controls, the Regulation governing accreditation[3] , the Decision governing conformity assessment mechanisms[4] and the Regulation on market surveillance[5] are indispensable.

With the AIA as part of the New Approach harmonisation model, I believe it is essential to stress at this point that this European standard has an extraordinary particularity compared to the rest of the legislative acts that follow this regulatory technique, since it serves as a legal basis for establishing harmonised standards for the whole of the Union, not for physical products, as has been traditional since the generalisation of the New Approach policy in the mid-1980s[6], but for the different categories of software that form part of the large family of Artificial Intelligence.

[2] Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council.

[3] Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation.

[4] Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products and repealing Council Decision 93/465/EEC.

[5] Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and product conformity and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011.

[6] In relation to the policy of the new harmonising approach see, for example, the books by Álvarez García, V., *Industria*, Iustel, 2010, pp. 47 ff, and *Las normas técnicas armonizadas (Una peculiar fuente del Derecho europeo)*, Iustel, 2020, pp. 21 ff; as well as the pioneering and essential works by M. López Escudero, *Los obstáculos técnicos al comercio en la Comunidad Económica Europea*, Universidad de Granada, 1991; by Valencia Martín, G., *La defensa frente al neoproteccionismo en la Comunidad Europea*, Cámara Oficial de Comercio, Industria y Navegación, 1993; and Mattera, A., *Le Marché Unique Européen, Ses régles, son fonctionnement*, Jupiter, 1988. The first of these works

2. A brief introduction to the basic elements of the new approach harmonisation technique as applied to Artificial Intelligence

An analysis of the harmonisation technique[7] of the new approach reveals that it has three main pillars: firstly, the existence of mandatory essential requirements that a product must comply with, which are directly established by the legislative act approved by the European Institutions (in the case of Artificial Intelligence, by a Regulation); secondly, the regulation of technical specifications that facilitate the justification of the product's conformity with these mandatory requirements; and thirdly, the existence of controls that ultimately make it possible to prove that the product complies with these mandatory essential requirements. Let us spend a few lines explaining these basic elements of the new approach as applied to Artificial Intelligence.

A) The overriding essential requirements to be met by high-risk Artificial Intelligence systems are aimed at the protection of health, safety, security, fundamental values and fundamental rights in the European Union.

These mandatory requirements set by the AIA are organised around the following areas: data and data governance, technical documentation, records, transparency and communication of information to users, human oversight, accuracy, robustness, and cybersecurity.

B) The technical specifications that enable compliance with the essential requirements imposed on Artificial Intelligence systems by their regulatory regulation are of three orders: firstly, European harmonised technical standards (or, more briefly, harmonised standards); secondly, common specifications; and thirdly, other technical solutions "at least equivalent" to the harmonised standards or common specifications referred to, and which may be provided, for example, by technical standards developed by standardisation bodies (international, European, and national) or by the economic operators behind the development of Artificial Intelligence software themselves.

C) Technical controls to demonstrate compliance with mandatory requirements for high-risk Artificial Intelligence systems are based on the idea that there is no point in setting standards if they are not accompanied by mechanisms to ensure compliance.

Under the new approach, product controls can be pre-market or pre-mar-

also contains an extensive bibliography on the historical construction and functioning of the European freedom of movement of goods, for those who wish to go more deeply into this question.

On this subject, I also consider it essential to consult the Commission's Communication entitled *Blue Guide to the implementation of EU product legislation in 2022* (OJEU C 247, p. 1 et seq.).

[7] See Álvarez García, V., *Las normas técnicas ...cit.* pp. 21 et seq.

keting, but can also be *ex-post*. This scheme is replicated in relation to high-risk Artificial Intelligence systems[8].

Ex-ante controls can be carried out either by the economic operator himself directly (this is the case of self-controls and self-certifications) or by an independent third party (which is a conformity assessment body previously notified to the European authorities – or, simply, a notified body) previously accredited by a body assigned to this task (notifying body).

Ex-post controls may also take different forms depending on the parties carrying them out. Indeed, these controls may be carried out by private operators (i.e., in the first instance, the supervision will be carried out by the providers themselves directly or by a third party, i.e., by the notified bodies), but also by public entities, given that the national market surveillance authorities play a fundamental role in the ultimate control of high-risk Artificial Intelligence systems marketed within the EU, without forgetting that the state authorities or bodies responsible for supervising or enforcing fundamental rights must be involved in this type of control.

Well, having focused on these three types of basic elements of the new approach harmonisation technique applied to Artificial Intelligence, over the next few pages I will focus exclusively on the second ones, meaning the two main categories of technical specifications that can be issued to develop the new Act governing this transcendental family of software that make up Artificial Intelligence, starting with the harmonised standards and continuing with the common specifications. In any case, it should be stressed before starting their individualised study that, unlike the essential requirements established directly by the new approach legislative act we are analysing (which are, let us recall, mandatory for high-risk Artificial Intelligence systems), both the harmonised standards and the common specifications are legally voluntary, although they are endowed with a legal-public effect of the first order: software generated according to these two types of technical documents is presumed to be in conformity with the aforementioned essential requirements that must be imperatively respected in order to be validly placed and marketed on the Community market[9]. This presumption of conformity opens up, in other words, this entire market to the product. In any case, it must be understood that this legal voluntariness means that economic operators can

[8] Álvarez García V. and Tahiri Moreno, J. "La regulación de la inteligencia ..." cit.

[9] On this public-legal effect of the presumption of conformity, see Álvarez García, V., *Las normas técnicas ...cit.* pp. 160 et seq. In relation to this issue, the conclusions of Advocate General Laila Medina presented on 22 June 2023 in the case "*Public.Resource.Org, Inc., Right to Know CLG v. European Commission*", C-588/21 P (points 33 et seq.) are very interesting.

follow alternative technical regulations to those laid down in the harmonised standards or in the common specifications in order to produce their products. However, this voluntary legal configuration has *de facto* clashed with the high bureaucratic and economic costs of manufacturing products according to other technical solutions, which increase the number of technical controls required to access the European market.

II. Harmonised standards

1. A basic preliminary question: harmonised rules have the legal nature of Community law

Since the new approach technique became widespread in the 1980s, harmonised standards have become a major instrument for the implementation of European legislation of this nature.

Despite the abundance of such technical documents and their great importance, it was only in 2016 that the Court of Justice of the European Union addressed their legal nature for the first time, declaring in its crucial *James Elliott* ruling that these standards constituted Union law[10].

The justification for this characterisation has been made by this High Community Institution on the basis of a double element: on the one hand, its elaboration process; and, on the other hand, the legal-public effect of the presumption of conformity to which I have already briefly referred.

With regard to their generation, these standards are drawn up by private subjects, which are the European standardisation bodies, following an internal procedure agreed within them. However, given their function of complementing legislative acts, the process of intervention by the Commission in their production is really very relevant: it is true that they are drawn up by standardisation bodies, but they do so at the request (or mandate) of this

[10] CJEU of 27 October 2016, case "*James Elliott Construction Limited v. Irish Asphalt Limited*", C-613/14. An extensive study of this judgment can be found at Álvarez García, V., "La confirmación por parte de la jurisprudencia del Tribunal de Justicia de la Unión Europea de la capacidad normativa de los sujetos privados y sus lagunas jurídicas (el asunto "James Elliott Construction Limited contra Irish Asphalt Limited")", *Revista General de Derecho Administrativo*, n.º 46, 2017. See also B. Lundqvist, "European Harmonised Standards as `Part of EU Law`: The implications of the James Elliott Case for Copyright Protection and, possibly, for EU Competition Law", *Legal Issues of Economic Integration*, no. 44, 2017; and A. Volpato, "The Harmonised Standards before the ECJ: James Elliott Construction", *Common Market Law Review*, no. 54(2), 2017.

High Institution; and once finalised by these bodies, they must be accepted by the Commission itself, which must also publish their references (i.e. their numerical code and title) in the Official Journal of the European Union, if they are to start to produce legal effects.

With regard to these legal effects, the Court of Justice of the EU recalls that, despite their voluntary nature, they enjoy a presumption of conformity, which means that they often become de facto mandatory, because, thanks to their compliance in the product manufacturing process, economic operators are greatly reduced in their burden of proving compliance with the essential requirements imposed by the legislative act (Directive or Regulation), which must be complied with in order to place and market the goods on the European market.

2. The distinction between European standards and harmonised European standards

Harmonised standards[11] are generally regulated in the European Standardisation Regulation 2012[12]. To this standard must be added, for each product subject to the harmonised scope of the new approach, the specific provisions that may be contained in its specific regulatory legislative Act. In the case of Artificial Intelligence systems, the particularities are really minimal, without really adding anything significant to the aforementioned general regulation of 2012.

Despite its name, the Regulation on European standardisation does not fully regulate this entire field, given that, although it contains a regulation on harmonised standards (and now also on standards implementing the General Product Safety legislation[13]), it does not perform this function with regard to simple European technical standards (or, in short, European standards).

These regulatory categories certainly have a common substrate: in both

[11] On this distinction, see, for example, the works of Álvarez García, V. "El lugar de las normas técnicas y de las normas técnicas armonizadas en el ordenamiento jurídico europeo", in Jiménez de Cisneros Cid, F.J. (Dir.), *Homenaje al Profesor Ángel Menéndez Rexach*, Aranzadi, 2018, pp. 99 et seq.; and *Las normas técnicas... cit.* pp. 71 et seq.

[12] Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council.

[13] Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on General Product Safety, amending Regulation (EU) No 1025/2012 of the European

cases they are technical specifications applicable to products (or services), which are drawn up by the European standardisation bodies and which are intended for repetitive or continuous application over time. In other words, they are really technical standards of a continental nature, originating from a private subject, following a private procedure and possessing a voluntary legal nature.

The differences between the two (i.e., between European technical standards and harmonised European technical standards) lie in the following three elements:

A) First of all, European standards are not intended to supplement any harmonising legislative act of the Union (they have an independent life), whereas harmonised standards are an essential development of the new approach legislative acts. It should be recalled that these legislative acts lay down the essential mandatory requirements that products must comply with in order to be validly placed and marketed on the European market, whereas harmonised standards lay down the technical specifications, compliance with which makes it possible to justify compliance by these products with these mandatory requirements. In this way, the New Approach harmonisation technique is based, as I indicated earlier, on a sort of public-private partnership in Europe, insofar as, on the one hand, the European institutions (public law entities) draw up the New Approach legislative acts that establish the essential requirements that products must comply with in order to be validly placed on the Community internal market and, on the other hand, these legislative acts are technically developed by the European standardisation bodies (private law entities) by means of the drafting of harmonised standards.

B) Secondly, harmonised standards are drawn up by the European standardisation bodies in accordance with procedures laid down by them, but the Community institutions are heavily involved in the drafting process. Thus, harmonised standards are used to implement new legislative acts adopted by the European Parliament and the Council; they are drawn up following a mandate from the Commission (it is true, however, that the procedure laid down by the standardisation bodies is followed in implementing the mandate, i.e., in drawing up the standard); once drawn up, they are examined by the

Parliament and of the Council and Directive (EU) 2020/1828 of the European Parliament and of the Council, and repealing Directive 2001/95/EC of the European Parliament and of the Council and Council Directive 87/357/EEC.

On this issue, see, from a doctrinal point of view, Álvarez García, V. "Los documentos técnicos normativos que sirven para garantizar la seguridad de los productos en la Unión Europea", *Revista General de Derecho Administrativo* n.º 64 Iustel, October 2023.

Commission and, if necessary, accepted by it; and finally, their references are published by the Commission in the European Official Journal. All this public intervention is non-existent in the case of purely European technical standards, which are drawn up freely by the European standardisation bodies in accordance with their internal procedures.

C) Thirdly, harmonised standards and European standards have, as we said a few moments ago, a voluntary legal nature, but harmonised standards are endowed with legal-public effects of the first order, which make them, as we already know, true acts of Community law auditable by the Court of Justice of the European Union. The main of these effects consists in the presumption of conformity of the product manufactured in accordance with its technical prescriptions with the mandatory requirements laid down by the new approach legislative act regulating it for the whole of the European internal market. It should be noted that harmonised standards produce these public-legal effects, which are essential to ensure the Community's freedom of movement of goods, despite the fact that they are drawn up by private parties and that their full content is not officially published, but only their references (i.e., their numerical code and title). Moreover, this content of the standards, which is translated into the official languages of the various Member States of the Union by their respective national standardisation bodies (thus transforming European standards into national standards), is protected by intellectual property rights belonging to the standardisation bodies, which sell them to finance themselves[14].

A final consideration with regard to the legal voluntariness of European standards and harmonised standards: this voluntariness means that, if economic operators do not follow them, they will not be punished administratively with any sanction, but the practical reality shows that the market often imposes them de facto, because economic operators tend to purchase only products manufactured in accordance with these standards and this is accredited by the corresponding certification[15]. To this market imposition, which is common to both categories of standards, is added in the case of harmonised standards their legal effect of presumption of conformity, which facilitates transnational trade within the Union and reduces the burdens (administrative

[14] On this issue, see Álvarez García, V., *Las normas técnicas... cit.* pp. 146 et seq.

[15] Regarding the question of the factual obligatory nature of both technical standards and voluntary certifications, see Álvarez García, V., "El proceso de privatización de la calidad y de la seguridad industrial y sus implicaciones desde el punto de vista de la competencia empresarial, *Revista de Administración Pública*, n.º 159, 2002, pp. 344 ff; y Muñoz Machado, S.*Tratado de Derecho Administrativo y Derecho Público General*, T. XIV: *La actividad regulatoria de la Administración*, BOE, 4th ed.

and economic) surrounding the demonstration that a product complies with the essential requirements of the new approach legislative act that regulates it.

3. The involvement of the European standardisation bodies and the Commission in the procedure for drawing up harmonised standards

The text of harmonised standards is drafted by one of the three European standardisation bodies, depending on the subject concerned. These three bodies, which have in common that they are private, associative bodies set up under Belgian or French law, are: 1) The European Committee for Standardisation (CEN), whose standardisation functions extend to all areas of industry and services (excluding electrotechnology and telecommunications); 2) The European Committee for Electrotechnical Standardisation (CENELEC), whose activity focuses on electrotechnology; and 3) The European Telecommunications Standards Institute (ETSI), which is responsible for drawing up standards in the world of telecommunications[16].

Harmonised standards are only drawn up by one or more of these European associations. It is true, however, that sometimes the continental bodies simply convert into European standards the standards adopted by their international counterparts, of which there are also three: the International Organisation for Standardisation (ISO), the International Electrotechnical Commission (IEC) and the International Telecommunications Union (ITU). It is true that the first two bodies, which are international non-governmental organisations (made up of the national standardisation bodies of most of the world's states), perform exclusively standardisation functions, while the ITU is a public-law body, as the United Nations specialised agency for information and communication technologies (ICT), and that, among the many functions it performs in this field, is the performance of standardisation tasks[17].

To the extent that the European standardisation bodies are private subjects, acting according to private procedures, their standards have traditionally been private. This is still the case for European technical standards. However, with the development of the new approach harmonisation technique, the European institutions have for decades been entrusting the abovementioned continental standardisation bodies with the task of drawing up technical solutions to complement the new approach legislative acts, which are known as harmonised standards.

This public-legal function that harmonised standards ultimately fulfil has

[16] Álvarez García, V., *La normalización industrial*, Tirant lo Blanch, 1999, pp. 367 et seq.

[17] *Ibid*, pp. 423 et seq.

traditionally justified, as we have already mentioned above, an intervention by the Commission in their adoption process, provided for specifically in each of the new approach legislative acts, but since 2012 there has been a general regulation in the Regulation on European standardisation ordering this intervention[18]. The essential milestones are:

A) The issuing of the specific standardisation mandate by the Commission, which is addressed to one or several European standardisation bodies and which, therefore, is prior to the start of the standardisation work of these bodies (which can always accept or reject it). In any case, without a prior mandate, there will be no harmonised standard. The essential regulatory provisions on mandates are set out in Article 10.1 and 10.2 of the 2012 Regulation, although the Commission has certainly developed their content extensively in its *Vademecum on European standardisation in support of European Union legislation and policies*[19].

B) The content of the harmonised standard is drawn up by the European standardisation bodies in accordance with their internal regulations. It is true that, during the drafting process, these bodies coordinate with the Commission, but the competence to draft the text of the standard and to approve it is exclusive to these bodies. When the standard is approved, the corresponding European standardisation body must send the Commission the full text of the standard in its official working languages (English, French, and German), together with the references of the standard (including its numerical code and title in all the official languages of the Union).

[18] On the intervention of the Commission in the process of drawing up harmonised standards, see Álvarez García, V., *Las normas técnicas... cit.* pp. 103 et seq.

[19] Within the limits of the powers provided for in the TFEU, the Commission may request one or more European standardisation organisations to draw up a European standard or a European standardisation deliverable within a specified time limit. European standards and European standardisation deliverables shall be market-based, take into account the public interest as well as the policy objectives clearly stated in the Commission's request, and be the result of consensus. The Commission shall set requirements for the content to be met by the requested document and a deadline for its adoption.

2. The decisions referred to in paragraph 1 shall be adopted in accordance with the procedure referred to in Article 22.3, after consulting the European standardisation organisations and European stakeholder organisations receiving Union financing in accordance with this Regulation, as well as the committee established by the relevant Union legislation, where such a committee exists, or through other means of consultation of sectoral experts".

The procedure referred to in Art. 10.2 of the European Standardisation Regulation 2012 is the examination procedure, which is regulated in Regulation (EU) No 182/2011 of the European Parliament and of the Council of 16 February 2011 laying down the rules and general principles concerning mechanisms for control by Member States of the Commission's exercise of implementing powers.

C) The Commission must analyse the content of the harmonised standard to verify whether its text is in line with the stipulations of the new approach legislative act that it implements and with the specific mandate that this Institution addressed to the standardisation bodies and which serves as a specific legal basis. This task of "reception" of the harmonised standard by the Commission is regulated in Article 10.5 of the Regulation on European standardisation[20].

Although it is not stated in the aforementioned precept, the process of verifying compliance with the mandate of the corresponding harmonised standard first involves consultants external to the Commission (or *Harmonised Standards Consultants* -HAS-) and, finally, the officials of this High Institution themselves[21].

D) In the case of acceptance of the harmonised standard by the Commission, the references (and only the references, not the text) shall be published in the Official Journal of the European Union. Only with this limited official publication, the harmonised standard will enjoy the public legal effect of presumption of conformity. The obligation to proceed to this official publication, once the harmonised standard has been taken over by the Commission, is foreseen in Article 10.6 of the Regulation on European Standardisation[22].

[20] This provision states that: "The European standardisation organisations shall report to the Commission on the activities undertaken for the preparation of the documents referred to in paragraph 1. The Commission, together with the European standardisation organisations, shall assess the conformity of the documents prepared by the European standardisation organisations with their initial request".

[21] On the process of reception of technical standards by the Commission, Álvarez García, V., *Las normas técnicas ...cit.* pp. 133 et seq.

[22] This provision foresees that: "Where a harmonised standard satisfies the requirements which it is intended to cover, laid down in the relevant Union harmonisation legislation, the Commission shall publish a reference to that harmonised standard without delay in the Official Journal of the European Union or by other means under the conditions laid down in the relevant act of Union harmonisation legislation".

Regarding the publicity of harmonised standards and the legal problems it raises, see Álvarez García, V., *Las normas técnicas ...cit.* pp. 136 et seq, and 182 et seq.; Bellis, M. De, "Private standards, EU law and access – The General Court's ruling in Public.Resource.Org", Eulawlive, 10-9-2021; Volpato, A. "Rules Behind Paywall: the Problem with References to International Standards in EU law", Eulawlive, 19-7-2021; Volpato, A., "Transparency and Legal Certainty of the References to International Standards in EU Law: Smoke Signals from Luxembourg: Stichting Rookpreventie Jeugd and Others (C-160/20)", Eulawlive, 1-3-2022; and Volpato A. and Eliantonio, M., "The Butterfly Effect of Publishing References to Harmonised Standards in the L series", European law blog, 7-3-2019.

4. Provisions on harmonised standards during the process of the Commission, the Council and the European Parliament during the procedure for the Proposal for a Regulation

The regulations on harmonised standards in the Commission's AIA proposal, the Council's compromise text and the Parliament's amendments contain a number of variations.

A) The provisions of the Commission's text contained in its brief Article 40 concern only the effect of the presumption of conformity with the mandatory essential requirements laid down in the AIA proposal for high-risk Artificial Intelligence systems complying with harmonised standards whose references have been published in the European Official Journal. It should be recalled that this fundamental public-legal effect causes, in the different products subjected to the harmonising legislation of the new approach within the European Union, that although the harmonised are legally voluntary, they are widely followed by manufacturers, since their compliance reduces bureaucratic and economic burdens and, in many cases, allows the use of the CE marking after mere internal conformity assessments (that is, carried out by the manufacturers themselves) and, therefore, access to the European internal market.

B) The Council's text also provides for this transcendental effect of the presumption of conformity, but adds some clarifications regarding, firstly, the content of the mandates that the Commission must issue to the European standardisation bodies, entrusting them with the drafting of harmonised standards implementing the provisions of the AIA in relation to high-risk Artificial Intelligence systems (to which it adds general-purpose Artificial Intelligence systems[23]), and, secondly, on the obligations of these standardisation bodies once they consider that this mandate has been fulfilled with the approval of the content of the corresponding harmonised standards.

a) The Council compromise text provides, in relation to the first of these questions, that standardisation mandates should specify that harmonised

[23] The concept of "general purpose Artificial Intelligence system" is incorporated in the Council compromise text as follows: it is, says this document, "an Artificial Intelligence system which, regardless of the manner in which it is brought to market or put into service, including open source software, has been designed by the provider to perform general purpose functions, such as image and speech recognition, audio and video generation, pattern detection, question answering, translation, etc. An Artificial Intelligence system can be used in a plurality of contexts and integrated into a plurality of other systems" [art. 3.1(b)] [art. 3.1(b)]. A general-purpose Artificial Intelligence system can be used in a plurality of contexts and integrated into a plurality of other systems" [Art. 3.1(b)].

standards should be "coherent" and "clear" and pursue, "in particular", these four objectives: firstly, ensuring that high-risk Artificial Intelligence systems are secure, respect EU values and guarantee "their open strategic autonomy"; secondly, promoting "investment and innovation in Artificial Intelligence, including by increasing legal certainty, as well as the competitiveness and growth of the EU market"; thirdly, promoting "investment and innovation in Artificial Intelligence, including by increasing legal certainty, as well as the competitiveness and growth of the EU market"; thirdly, the promotion of the participation ("governance") of all stakeholders in standardisation (from, for example, industry to civil society, SMEs and researchers); and fourthly, the strengthening of "global" cooperation in AI standardisation, in a way that is "consistent with EU values and interests".

b) Secondly, with regard to the obligations of the European standardisation bodies drafting harmonised standards, the Council in its compromise text lays down the burden on them to provide "evidence of their efforts to meet the objectives referred to".

C) The European Parliament, for its part, formulates four amendments to the short original text formulated by the Commission on the regulation of harmonised standards.

a) The first amendment (this is, 437) extends the effect of the presumption of conformity beyond high-risk Artificial Intelligence systems to foundational models[24].

b) The second and third amendments (i.e., 438 and 439) concern both the way in which requests (or mandates) are drawn up by the Commission and their content, with the following wording: 1) It empowers the Commission to formulate them in relation to all the essential requirements established by the AIA, respecting the provisions on this issue established by the Regulation on European standardisation; 2) It sets a maximum period of two months from the entry into force of the Regulation to submit the petitions; 3) In preparing the petitions, the Commission is required to consult the European Committee on Artificial Intelligence provided for in Article 56 of the Commission's proposal and finally in Article 64 and the consultative forum; 4) In formulating the content of these requests, the Commission is required to specify that

[24] The European Parliament's amendment 168 to the proposal for a Regulation on Artificial Intelligence proposes the introduction of a new point 1c to Art. 3 with the following definition of "foundational model": "a model of an Artificial Intelligence system trained on a large volume of data, designed to produce output information of a general nature and capable of adapting to a wide variety of different tasks". In the final version, the concept of "general-purpose AI model" is 3. 63rd handled.

the rules shall be "consistent" with the provisions of the Regulation and all harmonising legislation so far adopted within the Union, while reiterating the obligation that these rules are "aimed at ensuring that Artificial Intelligence systems or foundational models placed on the market or put into service in the Union comply with the relevant requirements set out in this Regulation" (both those set out for high-risk Artificial Intelligence systems and those set out for foundational models).

c) The fourth amendment (amendment 440) concerns the obligations incumbent on actors involved in the standardisation process. These obligations are the following four: 1) They shall take into account the general principles for trustworthy Artificial Intelligence as explicitly set out in the Regulation itself; 2) They shall seek to promote investment, innovation, competitiveness, and market growth, all in the field of Artificial Intelligence; (3) Contribute to the strengthening of international cooperation in the field of standardisation of Artificial Intelligence, also taking into account existing international standards in this field, provided that they are consistent "with the values, fundamental rights and interests of the Union"; and (4) Ensure the effective and balanced participation of all stakeholders in standardisation as referred to in the Regulation on European standardisation.

D) Finally, although the Commission's regulatory provisions on harmonised standards in the draft AIA are really brief, I do not believe that the additions proposed by either the Council or the European Parliament have either great declarative value or any practical effectiveness. It should be borne in mind that, as we have already seen a few moments ago, there is a general regime in the 2012 Regulation on European standardisation, which, although it is true that it is not very detailed, does establish a reduced legal regulation of the system for drafting and adopting harmonised standards, which is "completed" by other Commission documents such as the *Vademecum* to which I alluded earlier.

With this regulatory perspective, I believe that we should proceed to establish a regulatory framework for the European standardisation system as a whole, beyond the minimal regulation that is currently established for harmonised standards (and since a few months ago for the standards that develop the new General Product Safety Regulation[25]), which, as I said, is frankly limited, if not minimal. In any case, I am not going to focus now on the first of these issues (i.e., the establishment of a true general regulatory framework for the European standardisation system), but I would like to make a clarifica-

[25] Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on General Product Safety.

tion on the Commission's intervention in the process of developing harmonised standards, which I have insisted on on other occasions. This clarification is as follows: this type of technical standards implementing the new approach legislative acts have very important public-legal effects, and yet they are not published in the Community's Official Journal[26]. They are only published, in effect, in the Official Journal of the European Union. Only their numerical codes and titles are published, but not their content. The justification for their non-publication is that these texts, which are the property of the standardisation bodies, are used to finance them. It does not seem to me that the cost of these texts is so high that they cannot be purchased by the Commission[27], given that these standards are essential for the proper functioning of the harmonisation policy of the new approach (i.e., the internal market and European industry). In our country, we are aware of mixed public-private collaboration techniques that could be used to deepen relations between the Commission and the European standards organisations, which, moreover, already exist, being articulated through the General Guidelines for Cooperation between CEN, CENELEC, and ETSI with the European Commission and the European Free Trade Association of 28 March 2003[28].

The generation of harmonised standards is in fact a public task entrusted to private parties[29]. The Commission would have to compensate financially for the performance of these public tasks. This solution could contribute to improving the standardisation system by providing it with the necessary resources, which would be essential for all standardisation, and in particular for standardisation of Artificial Intelligence technologies. The latter stan-

[26] On the very important problem of the real lack of official publication of technical standards, see Álvarez García, V., *Las normas técnicas ...cit.* pp. 179 et seq.; and "La problemática de la publicidad oficial de las normas técnicas de origen privado que despliegan efectos jurídico-públicos", *Revista de Derecho Comunitario Europeo*, n.º 72, 2022. The conclusions of Advocate General Laila Medina, delivered on 22 June 2023, in the case "*Public.Resource.Org, Inc., Right to Know CLG v. European Commission*", C-588/21 P., are highly relevant to this issue.

[27] In this respect, please note the following from the Opinion of Advocate General Laila Medina delivered on 22 June 2023 in the case '*Public.Resource.Org, Inc, Right to Know CLG v European Commission*, C-588/21 P: "According to the CEN at the hearing, 4.6% of the standardisation budget comes from the sale of harmonised technical standards, which is equivalent to approximately EUR 2 million per year, whereas, in CEN's own words, the Commission's financing is equivalent to 'around 20% of CEN's total budget"' (point 99).

[28] On the content and significance of these General Guidelines, see Álvarez García, V., *Las normas técnicas ...cit.*, pp. 103 et seq.

[29] On the possibility of private parties drafting legal rules, see Álvarez García, V., "La capacidad normativa de los sujetos privados", *Revista Española de Derecho Administrativo*, n.º 99, 1998, pp. 343 ff; and *Las normas técnicas ...cit.* pp. 197 ff.

dardisation task requires sufficient financial and human resources to develop standards in short periods of time. Something that now, with the length of current standardisation processes, seems practically impossible. It should be borne in mind that "from the first proposal to final publication, the development of a technical standard usually takes three years"[30].

E) Regardless of the provisions on the legal effects of harmonised standards (common to the texts of the Commission, the Council, and the European Parliament), on standardisation mandates and the justification for compliance with them (specific to the texts of the Council and the European Parliament) or on the obligations of the actors involved in the standardisation process (specific to the Parliament), the three documents provide that the European Artificial Intelligence Committee (or, in the formulation of the European Parliament, the Office for Artificial Intelligence), in its role of advising and assisting the European Commission on this matter, may issue an opinion to the European Parliament, in the formulation of the European Parliament, all three documents provide that the European Committee on Artificial Intelligence (or, in the European Parliament's formulation, the Office for Artificial Intelligence), in its role of advising and assisting the European Commission on this matter, may issue "opinions, recommendations or written contributions" on existing harmonised standards or common specifications (Article 58(c) in the Commission's proposal and in the Council's text; and Article 56b(h)(i) of EP amendment 529). In what is finally Article 67.8th AIA.

5. Harmonised standards in the final text of the AI Act

The final version of the text of Article 40 of the AIA continues to regulate, unsurprisingly, the presumption of conformity of high-risk Artificial Intelligence systems that are in conformity with harmonised standards or parts thereof whose references have been published in the Official Journal of the European Union (in accordance with the provisions of the European Standardisation Regulation 2012) with the essential requirements applicable to them for their valid placing on the market within the Community territory.

Harmonised standards in the field of Artificial Intelligence, as in all other sectors covered by New Approach techniques, will be developed by the European standardisation bodies following a request or mandate from the Commission. The standardisation mandate, to be issued by the Commission "without undue delay", will require, firstly, results on information and doc-

[30] McFadden, M. et al. (Oxford Commission on AI&Good Governance), *Harmonising Artificial Intelligence: The role of standards in the EU AI Regulation*, December 2021, p. 17.

umentation processes to improve the resource efficiency of Artificial Intelligence systems and, secondly, specify that the standards must be consistent, clear and designed to ensure that Artificial Intelligence systems placed on the market or put into service within the Union meet the requirements set out in the AIA itself. In drawing up the standardisation mandates, the Commission will have to consult the Council and the various stakeholders (including the Advisory Forum).

The European standardisation bodies will develop the mandates by drawing up the relevant harmonised standards, according to their internal operating rules. However, they will have to provide evidence of their best efforts to meet the requirements and objectives set out in the AIA on request of the Commission.

Finally, the latter legislative act foresees that the actors involved in standardisation tasks should promote investment and innovation in the field of Artificial Intelligence. To this end, they shall seek to enhance legal certainty, competitiveness, and growth of the EU market, as well as contribute to strengthening global standardisation cooperation, taking into account existing international standards in the field of Artificial Intelligence, provided that they are consistent with the values, fundamental rights and interests of the EU, and improving multilateral governance with balanced representation of the interests involved and effective participation of stakeholders.

6. The problems of applying standardisation techniques to the Artificial Intelligence Act in the EU

Standardisation is a process of drawing up technical specifications which, reflecting the development of science and technology at a given moment in time, enable economic operators to manufacture their products[31]. These specifications, which in their most refined versions are called technical standards, are drawn up by standardisation bodies, which may operate at the general in-

[31] On the phenomenon of standardisation, see my following works: Álvarez García, V., *La normalización industrial... cit.*; *Industria*, Iustel, 2010; *Las normas técnicas ...cit.*; as well as Aubry, H., Brunet A. and Peraldi Leneuf, F., *La normalisation en France et dans l'Union Européenne*, Presses Universitaires d'Aix-Marseille, 2012; Bismuth, R., *La standardisation internationale privée (Aspects juridiques)*, Larcier, 2014; Cantero M. and Micklitz, M.W. (Eds.), *The Role of the EU in Transnational Legal Ordering: Standards, Contracts and Codes*, Edward Elgar Publishing, 2020; Carrillo Donaire, J.A., *El derecho de la seguridad y de la calidad industrial*, Marcial Pons, 2000; Contreras, J.L., *The Cambridge Handbook of Technical Standardization Law. Volume 2: Further Intersections of Public and Private Law*, Cambridge University Press, 2019; Delimatsis, P., *The Law, Economics and Politics of International Standardisation*, Cambridge University Press, 2015; J. Esteve Pardo, *Técnica, Riesgo*

ternational, supranational, regional or state level. In the Western world, these bodies usually have a private legal form, since they are made up mainly of representatives of economic operators and consumers, as well as academics and, increasingly, social organisations (for example, those defending the interests of workers or the environment), not forgetting the more or less intense participation of the various public administrations. Decisions on technical standards are, in any case, taken by consensus of the different subjects.

Standardisation has traditionally been used mainly in the world of physical products to ensure their interoperability, security, and quality. From physical products, it has been progressively extended to the world of services (although here its development is still limited) and is now being pursued in the field of Artificial Intelligence software. Parallel to this extension of the material scope of standardisation techniques, the same is being done with their purposes (beyond interoperability, industrial safety and quality) to pursue economic, social and political objectives of the first order. For example, in the field of Artificial Intelligence systems, standardisation processes in the EU, which are still in their infancy, aim to ensure that European industry plays a significant role at world level and that values and fundamental rights are respected on the old continent.

Standardisation bodies have tried to reflect scientific and technological developments in technical standards in a reasonable timeframe, but this, which has been relatively easy for physical products and even services, poses

y Derecho, Ariel Derecho, 1999; Falke, J., *Rechtliche Aspekte der Normung in den EG-Mitgliedstaaten und der EFTA*, Band 3: *Deutschland*, European Communities, 2000; G. Fernández Farreres, "Industria", in Martín-Retortillo Baquer S. (Dir.), *Derecho Administrativo Económico*, T. II, La Ley, 1991; F. Gambelli, *Aspects juridiques de la normalisation et de la réglementation technique européenne*, Eyrolles/Féderation des industries mécaniques, 1994; M. Izquierdo Carrasco, *La seguridad de los productos industriales*, Marcial Pons, 2000; Malaret García, E., "Una aproximación jurídica al sistema español de normalización de productos industriales", *Revista de Administración Pública*, n.º 116, 1988; R. Rodrigo Vallejo, "The Private Administrative Law of Technical Standardization", *Yearbook of European Law*, n.º 40, 2021; H. Schepel, and J. Falke, *Legal aspects of standardisation in the Member States of the EC and EFTA*, Vol. 1: *Comparative report*, European Communities, 2000; Schepel, H. and Falke J. (Ed.), *Legal aspects of standardisation in the Member States of the EC and EFTA*, Vol. 2: *Country reports*, European Communities, 2000; Schepel, H. *The Constitution of Private Governance. Products Standards in the Regulation of Integrating Markets*, Hart Publishing, Oxford, 2005; H. Schepel, "The new approach to the new approach: The juridification of harmonised standards in EU Law", *Maastricht Journal of European and Comparative Law*, n.º 20(4), 2013; Tarrés Vives, M. *Normas técnicas y ordenamiento jurídico*, Tirant lo Blanch, 2003; Van Gestel R. and Van Lochem, P. "Private Standards as a Replacement for Public Lawmaking?", in Cantero M. and Micklitz H.W. (eds.), *The Role of the EU in Transnational Legal Ordering*, Edward Elgar Publishing, 2020; and Van Waeyenberge, A. "La normalisation technique en Europe-L'empire (du droit) contreattaque", *Revue Internationale de Droit Économique*, n.º 32 (3), 2018.

major challenges in the world of information and communication technologies (ICT) in general[32], and Artificial Intelligence systems in particular, given the rapid advances in science and technology in these fields. December 2021 saw the publication of a major study on the strengths and, above all, the challenges for the European standardisation system in view of the adoption of the EU framework for Artificial Intelligence. This work is entitled *Harmonising Artificial Intelligence: The role of standards in the EU AI Regulation*[33]. It seems appropriate to highlight now some of the challenges facing standardisation in the face of this type of technology, which, it was stressed in that study, "is so new that standardisation bodies are only now beginning to draw up their plans for standardisation activity"[34]. Of these challenges, the following five are worth highlighting at this stage:

A) First of all, this study points out that the capacity (i.e., resources) of European standardisation bodies to deal with the development of harmonised technical standards in the field of Artificial Intelligence needs to be improved. There is a significant mismatch between the harmonised standards deemed necessary in the field of Artificial Intelligence and the technical specification developments that have actually taken place so far in the old continent[35]. This dysfunctionality is largely due to the lack of resources needed for standardisation, both in terms of the ever-present question of funding and in terms of the number of AI experts that need to be employed in standardisation work.

B) The second major challenge is the need to ensure meaningful participation in the development of harmonised European technical standards by the bodies in charge of protecting fundamental rights and public interests. It

[32] It is true, however, that the European Institutions have for years already produced *soft-law* documents advocating the need to develop standardisation in the field of ICT. Examples include the Commission's 2009 White Paper *on Modernising ICT standardisation in the EU – The way forward* (COM(2009) 324final) and the Commission's 2016 Communication *on ICT standardisation priorities for the digital single market* (COM(2016) 176final).

[33] McFadden, M. et al., *Harmonising... cit.* pp. 4 and 5.

In addition, numerous public and private works and documents have been produced in recent years to generate ideas on how to adapt the European standardisation system to the new challenges posed by information and communication technologies (ICT) in general, and Artificial Intelligence systems in particular. To name but a few: firstly, the *Rolling Plan for ITC Standardisation* from 2019 -the latest version is from 2023-; secondly, the *Bildt report on EU Standardisation* (2019); thirdly, the *Note from 17 member States to Council on Competitiveness* (2021); or fourthly, the public consultation carried out by the Commission on a new European standardisation strategy (the title of this new strategy is *Roadmap for a new European Standardisation Strategy* -June 2021-).

[34] M. McFadden et al., *Harmonising ...cit.* p. 4.

[35] *Ibid*, p. 18.

should be borne in mind in this respect that one of the main objectives of the AIA is to establish a set of mandatory requirements for high-risk Artificial Intelligence systems with the aim of minimising potential adverse effects on safety, health, and fundamental values and rights in the Union, and should therefore also be one of the main objectives of the harmonised technical standards. For this reason, it is clear that the harmonised European standards that serve to develop the mandatory essential requirements laid down in the AIA "will be more effective if they are drawn up in collaboration with experts in health, safety, and fundamental rights"[36]. The question is whether the European standardisation bodies are really ready, at the present time, to protect health, safety and fundamental values and rights in the Union. It certainly seems that there is a need to involve sectors (and therefore experts) within the standardisation bodies that have not traditionally been involved in standardisation work.

C) The third major challenge focuses on the need for harmonised standards to be "flexible enough to reflect the rapid evolution of Artificial Intelligence technology and products". There is currently "a mismatch between the speed of deployment of AI-based products and services and the development of technical standards"[37]. In this context, the process of acceptance and publication of harmonised standards by the Commission should be simplified and streamlined.

D) The fourth challenge is based on the need to strengthen cooperative relations between European and international standardisation bodies. In this area, two ideas must be taken as a starting point, which are sometimes difficult to reconcile: on the one hand, harmonised European standards must respect European values and fundamental rights, according to the AIA, which is not necessarily the case for international technical standards; and, on the other hand, Europeans have an interest in global, open, and interoperable technical standards that facilitate trade between the European Union and the rest of the world. Against this background, duplication of efforts between European and general international standardisation bodies should be avoided as far as possible, so that international standards are exploited within the European Union as far as possible, reducing the workload of European standardisation organisations to concentrate their activity in those areas where international standards do not exist, as well as helping to remove (or at least reduce) barriers to international trade. The need to avoid duplication between European and international standardisation is certainly not new. To this end,

[36] *Ibid*, p. 19.

[37] *Ibid*, p. 18.

the Vienna and Frankfurt agreements have been developed between, on the one hand, the European standardisation bodies (CEN and CENELEC) and, on the other, the international standardisation bodies (ISO and IEC). The study *Harmonising Artificial Intelligence: The role of standards in the EU AI Regulation* (p. 21) recalls that these agreements give priority to the standardisation work of the ISO/CEI conglomerate over that of CEN/CENELEC, but the AIA project "could change this in practice, putting CEN/CENELEC in the driving seat"[38], so that this European conglomerate would lead international standardisation while upholding European values and principles. In this respect, the study notes that among the various factors that could contribute to this outcome are, for example, the strong incentives for global economic operators to produce in compliance with harmonised European technical standards, as it would allow them to reduce the costs of conformity assessment in the European Union, thus having easier access to the large EU market[39].

E) The fifth challenge is the need to develop better tools for monitoring compliance with technical standards (in particular harmonised standards) in cooperation with standardisation experts, industry, and product providers.

7. Current results of standardisation work in Artificial Intelligence

While national, European and international standardisation is well developed in most of the physical manufacturing industry, this has not been the case for Artificial Intelligence despite the fact that this family of technologies has already reached a considerable use. The technical standardisation of Artificial Intelligence is still in its infancy.

Although there are various subjects capable of drawing up technical standards in the world of Artificial Intelligence, the only ones legally competent to generate harmonised standards are the European standardisation bodies. In this respect, Artificial Intelligence is no different from any of the physical products affected by the European harmonisation technique of the new approach.

Insofar as the AIA has just been approved, it has not yet been possible to issue specific harmonised standards for its development, although it is true that both ETSI and the CEN/CENELEC conglomerate have already been working on this issue for some months, as can be seen from the consultation of their standardisation agendas. The former is focusing primarily on the se-

[38] *Ibid*, p. 21.

[39] *Idem*.

curity area, while the latter is working more on the trust and ethics aspects[40]. To develop its standardisation work in this area, in 2018 ETSI created the *Industry Specification Group on Securing Artificial Intelligence* (ETSI ISG SAI), whose founding members included Telefónica[41]. In 2019, CEN/CENELEC set up the *Joint Technical Committee 21 on Artificial Intelligence* (JTC 21)[42].

However, the territorial level at which standardisation in Artificial Intelligence is currently most developed is the general international level, thanks to the work of the ISO/IEC conglomerate[43], which has the *Joint Technical Committee 1* (ISO/IEC JTC 1), highlighting within it, for our purposes, the Subcommittee *SC 42 on Artificial Intelligence*[44].

It should be noted that International standardisation is essential for European standardisation, as many of the CEN/CENELEC European technical standards are based on international standards developed by ISO/IEC[45]. The incorporation of ISO and IEC standards by their counterpart European standardisation bodies has been facilitated by the Vienna and Frankfurt agreements, which serve to organise the relations between all these standardisation bodies[46]. This close relationship between European and international standardisation makes it more than reasonable to assume that something similar

[40] *Ibid*, p. 10.

[41] It should be noted that this standardisation body is made up of more than 900 members from more than sixty countries. Unlike CEN and CENELEC, which are only made up of national standardisation bodies, ETSI membership is open to all those subjects or bodies interested in standardisation in the world of telecommunications. See, in this respect, Álvarez García, V., *La normalización industrial... cit.* pp. 367 ff.

[42] Among the various standard-setting texts on Artificial Intelligence on which this JTC 21 is working, the following two can be cited as examples: *prCEN/CLC/TR 17894 (WI=-JT021001) Artificial Intelligence Conformity Assessment*; or *prCEN/CLC/TR XXXX (WI=-JT021002) Artificial Intelligence: General Description of Artificial Intelligence Tasks and Functionalities related to Natural Language Processing*.

[43] IEC stands for International Electrotechnical Commission.

[44] Several technical standards have already been developed by ISO/IEC in this area. To give just a few examples, the following five are listed below: *ISO/IEC 22989 Artificial Intelligence – Concepts and terminology*; *ISO/IEC 23894 Information technology – Artificial Intelligence – Risk management*; *ISO/IEC 24668 Information technology – Artificial Intelligence – Process management framework for analytics using Big Data*; *ISO/IEC 38507 Information technology – IT governance – Governance implications of the use of Artificial Intelligence by organisations*; or *ISO/IEC 23053 Information technology – Artificial Intelligence – Assessment of classification performance of machine learning models*.

[45] McFadden, M. et al., *Harmonising... cit.* p. 31: "Of the approximately 3500 CEN/CENELEC technical standards cited in the OJEU, 44% are based on international standards".

[46] A brief note on the bases on which the relationship between international and European standardisation is based can be found in Álvarez García, V., *La normalización industrial... cit.* pp. 439 and ff.

could happen in the field of Artificial Intelligence. The problem is that while in the European Union, the AIA requires European standardisation bodies to develop harmonised standards with respect for European values and fundamental rights, this does not necessarily apply to general international bodies.

At the general international level, in addition to the international standardisation bodies (ISO and IEC), whose core work is the development of technical standards, there are other organisations of either a public or private nature that can also generate technical standards as an ancillary part of their core function. Among the entities of this type that produce technical specifications in the field of Artificial Intelligence, the following three can be mentioned: firstly, the ITU-T (*International Telecommunications Union-Telecommunication Standardisation Sector*); secondly, the IEEE (*Institute of Electrical and Electronics Engineers*); and thirdly, the W3C or *World Wide Web Consortium*.

Finally, since 2021, the European Commission has been monitoring the standardisation work on Artificial Intelligence carried out by all the above-mentioned bodies through its *AI Watch* service, which is reflected in its *Artificial Intelligence Standardisation Landscape* report[47] of that year, with a final version in 2023[48].

III. The common specifications

1. Some initial thoughts on its concept

Unlike harmonised standards, which already have a long tradition in the EU, common specifications[49] have a much more recent history (although, admittedly, with some isolated precedents[50]), as it is a legal instrument that has only been used with some semblance of generalisation since 2017 through two legislative acts: Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices; and Regulation (EU)

[47] Its full title is *AI Watch: Artificial Intelligence Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework*. Its authors are S. Nativi and S. De Nigris.

[48] This latest edition has been published under the title *AI Watch: Artificial Intelligence Standardisation Landscape Update*.

[49] On this type of regulatory instruments, see Álvarez García V. and Tahiri Moreno, J. "La regulación de la inteligencia ..." cit. and "Los instrumentos normativos ..." cit.

[50] It is true that Article 5(3) of Directive 98/79/EC of the European Parliament and of the Council of 27 October 1998 on *in vitro* diagnostic medical devices already provided for the adoption of this legal instrument. On the basis of this provision, the Commission even adopted a number of common technical specifications by its Decision 2002/364/EC of 7 May 2002.

2017/746 of the European Parliament and of the Council of 5 April 2017 on *in vitro* diagnostic medical devices[51].

This time difference means that we have been able to study the profile of harmonised standards acceptably well, explaining their advantages and also their serious legal problems[52], but we certainly know little about what are the contours and how common specifications are meant to operate[53].

The AIA contains the following definition of a common specification: it is, says Article 3.28 of this text, "a document, other than a standard, containing technical solutions proposing a way to meet certain requirements and obligations set out in this Regulation"[54] .

This definition allows us to understand only that common specifications are general provisions that, far from (harmonised) standards, set technical standards, which offer a way to comply with the essential requirements set in an imperative way by the new legislative act that these instruments are intended to develop.

This concept is, of course, rather vague. It does not answer, among many other basic questions, who should draw up this type of technical document, how it should be done, or what its effects are.

Well, in the absence of a general regulation (such as that which exists for harmonised standards in the European Standardisation Regulation 2012), the answers to all these questions must be sought in the regulation of this instrument in Article 41 AIA.

[51] Already in 2023, both Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery and Regulation (EU) 2023/1542 of the European Parliament and of the Council of 12 July 2023 on batteries and their waste, which regulates this regulatory instrument of the common specifications, have been adopted.

[52] Álvarez García, V., *Las normas técnicas ...cit.*

[53] Examples of the use of such implementing acts of Union law can be found in Commission Implementing Regulation (EU) 2020/1207 of 19 August 2020 laying down detailed rules for the implementation of Regulation (EU) 2017/745 of the European Parliament and of the Council as regards common specifications for the reprocessing of single-use devices; Commission Implementing Regulation (EU) 2022/2346 of 1 December 2022 laying down common specifications for the non-medical device groups listed in Annex XVI to Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices; and Commission Implementing Regulation (EU) 2022/1107 of 4 July 2022 laying down common specifications for certain *in vitro* diagnostic medical devices in Class D in accordance with Regulation (EU) 2017/746 of the European Parliament and of the Council.

[54] The definition is similar to the definition in Regulations 2017/745 and 2017/746. Thus, in Article 2 of both texts (paragraphs 71 and 74 respectively), common specifications are defined as "a set of technical or clinical requirements, other than a standard, which provides a means to fulfil the legal obligations applicable to a product, process or system".

2. The evolution of the regulation of the common specifications during the procedure of the Proposal for a regulation: from the Commission's draft to the European Parliament's amendments and the Council's compromise text.

A) The Commission proposal on AIA devoted its Article 41 to common specifications.

a) This precept grants, first of all, a very wide discretion to the Commission for the adoption of these normative instruments by foreseeing that this High Institution could generate them when there are no harmonised standards "or when the Commission considers that the relevant harmonised standards are insufficient or that there is a need to address specific issues related to security or fundamental rights".

b) Secondly, it stipulates that the common specifications shall have the status of implementing acts adopted by the examination procedure, in the processing of which it shall be obliged to seek the views of bodies or groups of experts.

c) Thirdly, these normative instruments are endowed with the public-legal effect of the presumption of conformity.

d) Fourthly, it is foreseen that providers may use alternative technical solutions to the common specifications, provided that they are at least equivalent to the common specifications.

B) The variants introduced by the Council compromise text to the regulation of the common specifications established by the AIA proposal.

The amendments proposed by the Council to the proposed regulation of common specifications are quite significant, aiming to "limit the Commission's discretion"[55], by specifying when the Commission could draw up this type of technical documents, how the procedure for their adoption will be developed or what the relations between harmonised standards and common

[55] Introduction to the Council compromise text, p. 7. This idea is complemented by the new wording of recital 65 of that compromise text, which underlines the exceptional nature that the adoption of common specifications should have in the following terms: "in the absence of relevant references to harmonised standards, the Commission should be able to establish, by means of implementing acts, common specifications for certain requirements provided for in this Regulation as an exceptional alternative solution to facilitate the provider's obligation to comply with the requirements of this Regulation, where the standardisation process is blocked or where there are delays in the establishment of an appropriate harmonised standard. If such delays are due to the technical complexity of the standard in question, the Commission should take this into account before considering the possibility of establishing common specifications".

specifications will be. In any case, it should be borne in mind that the regulation of common specifications provided for in the Council compromise text does not only concern high-risk Artificial Intelligence systems, but also those in general use.

a) With regard to the cases in which the Commission may issue common specifications, the Council proposes that this could only be done when, having previously requested the European standardisation bodies to draw up harmonised standards, the Commission's mandate has not been accepted by these bodies, when the standards have not been submitted by these bodies within the deadline set or, finally, when, having actually drawn up a standard, it does not comply with the mandate.

b) With regard to the procedure for their adoption, the Council envisages that it should be the examination procedure, as proposed by the Commission, but with the following conditions: firstly, the latter institution, when drawing up the specifications, must meet the same objectives which, according to the Council, must be required of the European standardisation bodies in the mandates issued by the Commission for the adoption of harmonised standards[56]; secondly, the Commission should seek the views of the relevant bodies or groups of experts; thirdly, the Commission should consult the European Committee for Standardisation after consulting the European Committee for Standardisation (ECSB); thirdly, the Commission should first consult the European Committee on Artificial Intelligence; and fourthly, the Commission should inform the committee provided for in Article 22 of the 2012 Standardisation Regulation (composed of representatives of the Member States and chaired by a representative of the Commission) that the requirements for adopting a common specification are met.

c) As regards the relationship between common specifications and harmonised standards, the Council provides for a sort of primacy of the latter over the former, stating that: "When the references of a harmonised standard

[56] It should be recalled that these objectives, set out in art. 40.2 of the Council compromise text, and which we have already referred to above, are: firstly, ensuring that high-risk Artificial Intelligence systems are secure, respect EU values and guarantee "their open strategic autonomy"; secondly, promoting "investment and innovation in Artificial Intelligence, including by increasing legal certainty, as well as the competitiveness and growth of the EU market"; thirdly, the promotion of the participation ("governance") of all stakeholders in standardisation (from, for example, industry to civil society, SMEs and researchers); and fourthly, the strengthening of "global" cooperation on AI standardisation, in a way that is "consistent with EU values and interests".

are published in the OJEU, the common specifications shall be superseded, as appropriate"[57].

d) Finally, the public-legal effect attributed to the common specifications by the Council text is that of a presumption of conformity. However, unlike the Commission's proposal for a regulation (which provided that if economic operators did not use the common specifications, they could validly use other technical solutions 'at least equivalent' to those set out in them), the Council's text is silent on this issue. In any case, I believe that there seems to be no doubt that the fact that it is not manufactured according to harmonised standards or common specifications does not prevent, from a legal point of view, the use of other technical solutions, even if it results in increased bureaucratic and economic burdens for providers introducing their Artificial Intelligence systems in the European internal market.

C) The European Parliament formulates several amendments to the Commission's text (from 442 to 448).

a) Among these amendments, those referring, firstly, to the limitation of the Commission's discretion when regulating the cases in which the adoption of common specifications is appropriate stand out, since it is required that there is no prior harmonised standard and that the European standardisation bodies have been previously requested to adopt them (and they do not fulfil this task adequately). The Commission must also justify the reasons why it has decided to use common specifications.

(b) Secondly, the Commission shall, prior to its adoption, also consult the Office for Artificial Intelligence and the Consultation Forum, the European standardisation bodies, expert groups established under sectoral Union legislation and other relevant parties. Where the Commission decides not to follow the opinion issued by the Office for Artificial Intelligence, it shall provide a reasoned explanation.

c) Thirdly, the Commission will have to justify the fulfilment of the same objectives that the European Parliament wants to be required for the development of harmonised standards: (1) It shall take into account the general principles established by the Act for reliable Artificial Intelligence; (2) It shall seek to promote investment, innovation, competitiveness, and growth of the Artificial Intelligence market at Union level; (3) It shall contribute to strengthening global cooperation on standardisation, taking into account international standards on Artificial Intelligence, where they are "consistent with the values, fundamental rights and interests of the Union"; and (4) It shall ensure a

[57] Art. 41.4 of the Council compromise text.

balanced and effective participation of all stakeholders in the field of Artificial Intelligence standardisation.

d) Fourthly, the European Parliament proposes the following rule for resolving conflicts between harmonised standards and common specifications: "When the reference of a harmonised standard is published in the Official Journal of the European Union, the Commission shall repeal the implementing acts" adopting the common specifications or parts thereof, insofar as they cover the same subject matter.

e) Fifthly, the provisions on common specifications apply both to those developing the essential requirements for high-risk Artificial Intelligence systems and those for foundational models.

3. The essential elements that make up the common specifications in the AI Act

A) The Commission is responsible for drawing up and adopting these instruments. Unlike harmonised standards, the common specifications are entirely public technical documents, since they are drawn up entirely by the Commission, which takes the initiative, drafts the text and adopts it. It should be recalled that harmonised standards were generated at the initiative (mandate) of the Commission, but their content was drawn up by the European standardisation bodies, although it depended on the Commission's acceptance and the official publication of their references as to whether these standards could enjoy the public legal effect of presumption of conformity.

B) In which cases could common specifications be adopted? The handling of the AIA proposal revealed important divergences of appreciation in deciding when a common specification should be issued between the Commission on the one hand and the Council and the European Parliament on the other.

In contrast to the Commission's argument that it has a wide discretion to issue such technical documents whenever it sees fit (regardless, for example, of whether harmonised standards already exist on the specific issue), the final text of Article 41 AIA restricts this possibility by requiring the following three conditions to be met: (1) that a harmonised standard has not already been developed; (2) that the Commission has already issued the appropriate standardisation mandate to one or more European standardisation bodies; and (3) that, alternatively, these bodies have not accepted such a request, or there are undue delays in the adoption of the harmonised standard, or the harmonised standard does not comply with the Commission's mandate.

C) The drafting procedure is the examination procedure. This examination procedure is governed by Article 5 of Regulation (EU) No 182/2011 of the European Parliament and of the Council of 16 February 2011 laying down the rules and general principles concerning mechanisms for control by Member States of the Commission's exercise of implementing powers. This European legislative act requires the Commission, prior to the adoption of the relevant implementing act, to obtain a favourable opinion on its draft from the relevant committee composed of representatives of the Member States and the Commission itself (which chairs the committee, but has no vote). The majority required for such an opinion is the qualified majority provided for in Article 238.3 TFEU.

In the course of the procedure for drawing up the common specifications, the Commission shall be obliged to consult various experts in the fields of Artificial Intelligence and standardisation. In this regard, consultation of the Office for Artificial Intelligence and the Advisory Forum, European standardisation bodies, expert groups established under relevant sectoral Union law and other interested parties shall be required.

(D) the legal form to be taken by the common specifications

This legal form will be that of a Commission implementing act. It should be recalled in this regard that the legal basis for implementing acts of European Union law is Article 291 TFEU. This provision lays down, as a general rule, that measures implementing legally binding acts of the Union are a matter for the Member States, which shall adopt all necessary measures of national law to that end (paragraph 1). However, in cases which "require uniform conditions for implementing legally binding Union acts, implementing powers are conferred on the Commission" (or, in certain limited cases, on the Council) (paragraph 2), to be exercised, on the one hand, by the European institutions and, on the other hand, subject to the arrangements for control by the Member States, in accordance with the rules laid down in European regulations adopted under the ordinary legislative procedure (paragraph 3). The secondary legislation currently regulating this issue is the aforementioned Regulation (EU) No 182/2011.

These Commission implementing acts (and therefore the common specifications) must be published in full in the Official Journal of the European Union (Article 297(2) TFEU). It should be recalled that this situation is quite different from that of harmonised standards, which, although they have the presumption of conformity as a public-legal effect (as is the case for common specifications), have official publication restricted to their references (i.e., their numerical codes and titles).

E) The legal value and effects of technical specifications.

The general rule that seems to follow from the AIA is that these regulatory instruments are voluntary from a legal perspective. Let me explain: economic operators wishing to place on the market or make available on the market an Artificial Intelligence software must respect the essential requirements imperatively established by their new approach regulatory legislative act, and these mandatory requirements can be made by following either the harmonised standards that may exist (elaborated by the European standardisation bodies), either the common specifications (produced by the Commission), or other technical solutions (established, for example, by a private operator or by a national standardisation body – such as UNE in Spain – or a general international standardisation body – such as ISO) which can provide a level of quality and safety at least equivalent to that laid down in the existing common specifications in this field. In other words, the common specifications offer technical solutions that have alternatives that can be freely chosen from a legal point of view by each economic operator.

The decision to develop an Artificial Intelligence system in accordance with the common specifications, as with harmonised standards, carries a unique legal effect. In fact, when economic operators use the common specifications, there is a legal presumption that the systems made according to those specifications meet the essential requirements set by the relevant new approach legislative act. These requirements must be met in order for the systems to be placed in the European market.

In short, despite the public-legal character linked to the competent authority for the production of the common specifications and the procedure used to draw them up, this legal instrument is not legally binding, but is "only" conferred the same public-legal effect as harmonised standards, to which we have just referred, i.e., its presumption of conformity. This means that, despite the existence of these common specifications and/or harmonised standards, Artificial Intelligence operators may adopt alternative technical solutions for the development of their systems, provided that they can prove that they have "adopted technical solutions at least equivalent" to those laid down in the harmonised standards and common specifications.

G) The legal relationship between common specifications and harmonised standards.

We have seen that the Commission adopts common specifications when there are no technical standards or when, if there are technical standards, they are insufficient to regulate a subject. In other words, common specifications serve to fill the gaps left by the European standardisation bodies.

It does not seem sensible that when common specifications are adopted on a particular issue, there should be harmonised standards regulating that issue. The Commission has the power either not to publish in the Community

Official Journal the references of harmonised standards drawn up by the European standardisation bodies if it does not agree with them or to withdraw such references from the Official Journal if they are already published[58]. In other words, the Commission holds the key to avoiding this type of potential conflict.

In any case, it seems that a rule should be established to resolve potential conflicts between the two regulatory categories. The AIA appears to accomplish this. This legislative act provides for "a sort of primacy" (or even "hierarchy") of technical standards over common specifications, by stipulating that: "When the reference of a harmonised standard is published in the Official Journal of the European Union, the Commission shall repeal" the common specifications. In other words, the repeal of common specifications (public law standards) by harmonised standards (standards of private origin) is not automatic but only imposes on the Commission the obligation to repeal common specifications that are contrary to harmonised standards.

But, in the light of this text, doubts remain: what would happen in the reverse situation (if common specifications are adopted after harmonised standards)? Is there a hierarchical relationship between harmonised standards and common specifications? The answers to these questions are really uncertain, and can only be deduced through legal common sense, since we have neither normative nor jurisprudential elements that would allow us to undertake such a task.

H) The need for a general regulation of common specifications.

New approach legislative acts have traditionally been developed by harmonised standards, but these now seem to have found inseparable companions in the form of common specifications. Since the regulation of medical devices in 2017, and especially with the recent Machinery and Battery Reg-

[58] Article 11.1 of the European Standardisation Regulation 2012 (as amended by Article 48 of the General Product Safety Regulation 2023) provides for this possibility. Where a Member State or the European Parliament considers that a harmonised standard or a European standard developed in support of Regulation (EU) 2023/988 does not entirely satisfy the requirements that it is intended to cover, as set out in the applicable Union harmonisation legislation or in that Regulation, it shall inform the Commission thereof with a detailed explanation. After consulting the committee established by the relevant Union harmonisation legislation, where such a committee exists, or the committee established by that Regulation, or after other forms of consultation of sectoral experts, the Commission shall decide to: (a) to publish, not to publish or to publish with restriction the references of the harmonised standard or European standard concerned drawn up in support of that Regulation in the Official Journal of the European Union; and (b) to maintain or to maintain with restriction the references of the harmonised standard or European standard concerned drawn up in support of that Regulation in the Official Journal of the European Union or to withdraw them from it.".

ulations (both 2023), the essential requirements that products must fulfil to be validly placed and marketed on the European market can be completed by one or the other instrument.

It seems certain, therefore, that common specifications are a type of technical provisions that are here to stay, since they undoubtedly offer an important advantage from a practical perspective. At least on paper, these specifications would make it possible to fill the gaps that are not filled by the European standardisation bodies[59], when they are not in a position to adopt a harmonised standard in line with the needs of the Union (for example, when they are not able to meet the requirements arising from the Union's values or the protection of human rights).

However, if the European legislator decided to adopt a cross-cutting act establishing a general framework for harmonised standards (I am referring, of course, to the 2012 Regulation on European standardisation), which has made it possible to resolve important legal questions on their drafting and implementation, it seems to me that the same should be done for common specifications. I believe that the drafting of a general regulation on common specifications should be tackled, since, when analysing the few provisions on them in the recent new legislative acts that provide for them, their profiles are sometimes extremely vague, raising many legal doubts at both the drafting and implementation stages.

4. Other technical solutions equivalent to those offered by harmonised standards and by the common specifications

Under the new approach, harmonised standards generated by European standardisation bodies have traditionally not been the only technical solutions available to economic operators for the legal manufacture of their products, as they have always been offered the possibility of using alternative technical means to do so. These means include, for example, the use of international technical standards, non-harmonised European standards, national standards or simply technical specifications established by the economic operators themselves.

This legal configuration of voluntariness has, however, *de facto* clashed with the bureaucratic and economic costs that controls of products manufactured by the use of other technical mechanisms have had. In other words, although

[59] In this respect, it has been stated by leading experts in the world of standardisation that: "Common specifications act as a safety net or a safety barrier, empowering the Commission to act when there is a gap in the technical standards space" [McFadden, M. et al.]

harmonised standards are legally voluntary in use, practice shows that this is the system that economic operators prefer to use to manufacture their products, given the advantages it brings, linked to the presumption of conformity[60].

This situation has not changed with the introduction of the common specifications developed by the Commission. Both these and the harmonised standards are legally voluntary, allowing the use of other technical solutions "at least equivalent" to them. This is provided for in the AIA. The proposal for this legislative act presented by the Commission in April 2021 specified, in this respect, that these alternative technical solutions can be "developed on the basis of general scientific or engineering knowledge, at the discretion of the provider of the Artificial Intelligence system concerned. This flexibility is of particular importance, as it allows providers of Artificial Intelligence systems to decide how they want to meet the requirements, taking into account the state of the art and technological and scientific developments in this field"[61]. It remains to be seen whether in the near future standardisation will be a sufficiently agile mechanism to establish harmonised regulatory standards for Artificial Intelligence systems, and whether, failing that, the Commission will be able to fill the gaps left by the European standardisation bodies; or whether, on the contrary, economic operators will be more effective in establishing the technical specifications necessary for the development of a field that is developing as rapidly as Artificial Intelligence is evidently doing.

It should be recalled, in any case, and finally, that when these alternative technical solutions to harmonised standards or common specifications are used, there will be no presumption of conformity of the Artificial Intelligence systems generated in accordance with them with the mandatory requirements established by the AIA, thus significantly multiplying the difficulties and costs of this demonstration for the providers of these systems.

[60] In this regard, please note the following in the Opinion of Advocate General Laila Medina delivered on 22 June 2023, in the case '*Public.Resource.Org, Inc, Right to Know CLG v European Commission*, C-588/21 P: "The fact that harmonised technical standards are *de facto* mandatory, as they are often the only accepted method in the market to ensure compliance with the relevant Union secondary legislation, is confirmed by a study commissioned by the Commission: 'in practice, [harmonised technical standards] are almost mandatory for most economic operators'. Moreover, the same study points out that the price of harmonised technical standards is one of the main obstacles to their effective use' (point 45). This Commission study to which the Advocate General refers is made explicit in footnote 24 of her Opinion: *EIM Business & Policy Research, Access to Standardisation – Study for the European Commission, [DG] Enterprise and Industry*, 2010.

[61] Explanatory Memorandum to the Commission's proposal for a Regulation on Artificial Intelligence, p. 16.

CONFORMITY ASSESSMENT IN THE DESIGN AND PRODUCTION OF ARTIFICIAL INTELLIGENCE-BASED SYSTEMS IN THE CONTEXT OF THE "NEW LEGISLATIVE FRAMEWORK"

Adrián Palma Ortigosa

Lecturer in the Department of Administrative Law of the Universitat de València[1]

I. Introduction

This paper studies the content of the AIA that regulates the conformity assessment process for AI systems[2]. Firstly, an approach is made to the legal framework of the conformity assessment process within European legislation, which contemplates a whole series of instruments that seek to guarantee that certain products can be marketed in the European Union with certain guarantees and homogeneous requirements. Secondly, the conformity assessment mechanisms designed by the AIA for the different AI systems are analysed. Depending on the type of AI system, one or another procedure for verifying compliance with the requirements for AI systems will be applied.

[1] This work has been carried out in the framework of the following research projects:

"Algorithmical Law" (PROMETEO/2021/009. Funded by the Generalitat Valenciana. "Regulation of the digital economy: public protection of equality and algorithmic tools" (PID2019-108745GB-I00). Ministry of Science and Innovation. "Public rights and guarantees in the face of automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by the Ministry of Science and Innovation. "Algorithmic Tools for Citizens and Public Administrations" (Proyectos de Generación de Conocimiento, Ministerio de ciencia e Innovación, convocatoAIA 2021, PID2021-126881OB-I00). "Algorithmic Decisions and the Law: Opening the Black Box" (TED2021-131472A-I00) of the Recovery, Transformation and Resilience Plan. Digital Rights Agreement-SEDIA Scope 5 and 6 (2023/C046/00228673).

[2] This work has been carried out in the framework of the following research projects: "Algorithmical Law" (PROMETEO/2021/009. Funded by the Generalitat Valenciana. "Regulation of the digital economy: public guardianship of equality and algorithmic tools" (PID2019-108745GB-I00). Ministry of Science and Innovation. "Public rights and guarantees in the face of automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by the Ministry of Science and Innovation. "Algorithmic tools for citizens and Public Administrations" (Proyectos de Generación de Conocimiento, Ministerio de ciencia e Innovación, convocatoAIA 2021, PID2021-126881OB-I00).

II. Evolution, processing, and final content of the articles of the Artificial Intelligence Act involved

The articles governing the conformity assessment process in the AIA are as follows:

Recital	Precept	Subject
	Article 16(f)	Obligation for providers to ensure that their system undergoes conformity assessment before it is placed on the market.
Recital 122	Articles 41.3 and 42	Presumptions of conformity.
Recitals 123-125 and 128	Article 43	Conformity assessment foreseen for each type of AI system.
Recital 130	Article 46	Compliance exemptions.
	Annex VI	Conformity assessment carried out by the provider itself.
	Annex VII	Conformity assessment carried out by third parties.

These articles have remained largely unchanged since the initial proposal was adopted by the European Commission on 21 April 2021[3].

It should be noted that initially conformity assessment was not only regulated in the articles listed in the table above, as it was also mentioned as an obligation for providers in Article 19[4]. The content of this article was deleted during the processing of the AIA. This change has had no practical significance, as the original Article 19 on conformity assessment, now deleted, referred to the conformity assessment process regulated in Article 43 and Annexes VI and VII, which have remained virtually unchanged since the original version. We understand that this suppression is essentially due to two reasons: on the one hand, the fact that it was a perfectly dispensable reference, and, on the other hand, because the same obligation was also mentioned in Article 16.e of the initial text of the AIA[5].

[3] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislation. 21 April 2021.

[4] Former Article 19 of the original AIA proposal set out the obligation for the provider to ensure that the conformity assessment referred to in Article 43 was carried out.

[5] This is currently regulated in Article 16(f) of the AIA, which states that providers "*ensure*

III. Conformity assessment in European Union legislation

The conformity assessment process envisaged in the AIA and which will be detailed below, follows the scheme established by the so-called "New Legislative Framework", hereinafter referred to as the NLF. The NLF is made up of several European legal texts that establish a common basis for the marketing, assessment, and surveillance of products in the European Union[6]. Consequently, the European legislator, when legislating on a product, can take as a reference the NLF[7], which contemplates a structure that seeks to ensure a reliable evaluation and placing on the market of such products and goods.

The EU's approach to regulating the design, marketing, and oversight of AI systems is to maintain the NLF structure that has been in place for years for a number of other products[8].

In summary, and following the structure set out by the NLF, in order for a product to be marketed in the EU with minimum safety guarantees, the laws regulating these products must establish the following elements.

Firstly, products must comply with a number of *minimum technical requirements* that manufacturers have to implement. These essential requirements aim to mitigate the main risks that these products may cause once they are placed on the market. In the case of the AIA, these requirements are found in articles 8 to 15, including data quality, appropriate levels of transparency, accuracy, or robustness metrics, etc. The aim is that all appropriate technical measures are implemented in the design to mitigate the perverse effects that may affect the fundamental rights, safety, or health of individuals once the AI system starts to be used.

Secondly, there is the possibility for manufacturers to use certain technical standards developed by European standardisation bodies[9] or by the European

that the high-risk AI system undergoes the relevant conformity assessment procedure as referred to in Article 43, prior to its being placed on the market or put into service".

[6] The three legal texts that make up the New Legislative Framework are: Regulation (EC) No 765/2008 of the European Parliament and of the Council setting out the requirements for accreditation and market surveillance of products; Decision No 768/2008/EC of the European Parliament and of the Council on a common framework for the marketing of products and; Regulation (EU) 2019/1020 of the European Parliament and of the Council on market surveillance and product conformity.

[7] In addition to the New Legislative Framework, the so-called New Approach and the Global Approach must also be taken into account. A historical analysis of these can be found in: V. Álvarez García, *IndustAIA*, Iustel, 2010, pp. 47 et seq. See also the Blue Guide on the implementation of the European Product Regulations 2022. pp.7 et seq.

[8] This is expressly stated in different recitals of the AIA. See recitals 46, 64, 83, 84, 87, 124.

[9] These European standardisation bodies are CEN, CENELEC and ETSI.

Commission that are specifically designed to meet the technical requirements of the legislation for that product. We are referring to *harmonised standards and common specifications* respectively[10]. These standards are discussed in another chapter of this collective work.

Thirdly, once the manufacturer has implemented the requirements, whether or not by reference to harmonised standards or common specifications, the next step is for the product to undergo a *conformity assessment* to ensure that it complies with the legal requirements set out in the legislation. Each product has its own type of conformity assessment, although there are some similar conformity assessment processes with their own specifications applicable to each product[11]. In some cases, this assessment will be carried out by a third party other than the manufacturer.

Fourthly, once the product has passed the conformity assessment, it is up to the manufacturer to *declare the conformity* of the product and, where appropriate, to establish the *marking* of the product. It is then that the product can be placed on the market.

Finally, fifthly, once the product is placed on the market, the manufacturer is obliged to continue to comply with the technical requirements of the legislation applicable to the product. In addition, the relevant public authorities will have the power to monitor and supervise that this is indeed the case, this is called *market surveillance*[12].

New Common Legislative Framework	**European AI Act**
Minimum compliance with requirements	Articles 8-15
Implementation of harmonised standards/ common specifications	Articles 40 and 41
Conformity assessment	Article 43
Declaration of conformity and CE marking	Articles 47 and 48
Market Surveillance	Articles 70 and 74 to 84

Products such as machines, toys, lifts, medical devices, and others follow the structure previously indicated[13], which is also present in the AIA for AI systems.

[10] See inter alia Articles 40 and 41 of the AI Act, as well as the definitions given in the Blue Guide on the implementation of the 2022 EU product legislation. P.49.

[11] Blue Guide on the implementation of the European 2022 product legislation. pp.74 ff.

[12] Market surveillance aims to ensure that products comply with applicable requirements that provide a high level of protection of public interests protected by EU harmonisation legislation.

[13] The full list of products and product safety components is mentioned in recital 50 of the AIA.

IV. The forms of conformity assessment in the Artificial Intelligence Act

Conformity assessment is the process of demonstrating that a product meets the requirements specified in a norm or standard[14]. In our case, through conformity assessment, providers or an external body verify that an AI system complies with the minimum requirements of the AIA[15].

The conformity assessment procedure is made up of a series of processes and phases through which it is verified that a product is in conformity with the requirements of the harmonisation legislation required for such a product to be placed on the market with certain guarantees. In this sense, if the AI system undergoes a substantial modification[16], it will be necessary to resubmit it to a new verification of conformity process[17].

The AIA provides for two main conformity assessment processes. The essential difference is who is responsible for verifying that the AI system complies with the AIA requirements before it is placed on the market. Thus, either the conformity assessment is carried out by the provider who has designed the system, or it is carried out by a third party, the so-called notified body. Each of these processes is described in Annexes VI and VII respectively.

It is not up to the providers to choose one or the other conformity assessment procedure but will depend on the type of AI system they have developed[18].

It is now time to look at each of the conformity assessment processes covered by the AI Act.

1. Conformity assessment procedure based on internal control

Through self-assessment of conformity or internal control, the provider verifies that its AI system complies with the requirements of the AIA. This

[14] ISO/IEC 17000:2004. Conformity assessment – Vocabulary and general principles.

[15] Conformity assessment means *the process of demonstrating whether the requirements laid down in Chapter II, Section 2 have been fulfilled in relation to a high risk AI system*. Article 3.20. AIA.

[16] A substantial modification in this context will be for example a change of operating system, a change in software architecture, a change in the intended purpose of the system, etc. Recital 128.

[17] Article 43.4 of the AI Regulation.

[18] Article 43 of the AI Regulation describes which type of conformity assessment each AI system falling under the scope of the AI Regulation has to undergo.

whole process is entirely carried out by the provider without intervention of third party notified bodies or public authorities.

According to Annex VI, the conformity assessment procedure based on internal control consists of three phases or processes.

Firstly, the provider has to verify that the implemented quality management system complies with all the requirements of the AIA. Thus, article 17 of this standard obliges the provider to develop a set of documentation and implement procedures to ensure that the quality management system is indeed adequate.

Secondly, it is up to the provider to assess that the AI system complies with the essential requirements foreseen by the AIA by reference to the technical documentation that has been prepared on the product[19]. The assessment of the requirements will oblige the provider to deploy different evaluation measures such as documentation checks, tests, testing of the AI system, etc. We understand that this whole process should be properly documented.

Thirdly, as a final step of the self-assessment process, the provider shall verify that the design process and the post-market surveillance of the AI system referred to in Article 72 of the AIA are consistent with the part of the technical documentation referring to these processes[20].

Once the provider has carried out these three processes in an optimal way, the conformity assessment is deemed to have been passed. Of course, all these internal actions should be documented and always be available to the market surveillance authority requesting such information.

The documentation of the entire conformity assessment procedure based on internal control is very important, as it demonstrates that the self-assessment has been carried out properly and confirms that the AI system complies with the requirements of the AIA.

2. The assessment carried out by a notified body

Alongside self-assessment, the other conformity assessment process foreseen by the AIA is the one involving a notified body. A notified body is an entity that has been accredited to carry out conformity assessments on AI systems subject to AIA and notified as such to the European Commission[21].

Annex VII of the AIA sets out the stages of the notified body's involve-

[19] Further information on the content and essential elements of the technical documentation can be found in Article 11 and Annex IV of the AI Regulation.

[20] See paragraphs 2 and 9 of Annex IV of the AI Regulation.

[21] See the work in this collective work which analyses the role and functions of notified bodies and, where appropAIAte, notifying authorities.

ment in the verification process of AI systems. This assessment essentially comprises the examination of the quality management system and the technical documentation of the AI system[22].

2.1.The assessment of the quality management system

As regards the management of the quality system, the provider must lodge an application for assessment of the AI system with the notified body. The contents of the application must include the provider's identification data, the technical documentation of the AI system developed, the documentation concerning the quality management system, as well as the procedures in place to ensure that the quality system will be complied with. We understand that each notified body will have different application forms or application forms to be submitted by the providers[23].

Once the application has been submitted, it is the responsibility of the notified body to assess whether or not the system complies with the requirements of Article 17 of the AIA. This decision must be notified to the provider or, where applicable, to his authorised representative. The decision must state the reasons for the decision and include the conclusions of the assessment.

Once the quality management system has been approved, changes may be made to the quality management system, in which case the provider shall, before making such changes, inform the notified body so that it can examine the proposed changes and decide whether they still meet the requirements laid down by the AIA. The notified body shall communicate its decision to the provider. This decision shall include the conclusions of the examination of the changes and the reasoned assessment decision.

In addition to deciding on potential changes that the provider intends to make to the quality management system, the notified body may carry out various surveillance and control activities in relation to the quality management system. Among others, the notified body is authorised to enter the provider's premises, to carry out periodic audits and to perform any additional tests it deems necessary to ensure that the AI system complies with the AIA.

[22] The European Commission is empowered to amend any of these steps by means of a delegated act. Article 43.5 AI Regulation.

[23] The National Centre for the Certification of Medical Devices is the only notified body in Spain to carry out the conformity assessment of medical devices in accordance with Regulation 2017/745. The application for the verification of the quality management system can be found on the following website:

https://certificaps.gob.es/wp-content/uploads/CertificacionMDR/R_DEX_05-Solicitud-de-evaluaci%C3%B3n-del-sistema-de-gesti%C3%B3n-de-calidad.pdf

2.2. Analysis of technical documentation

As was the case for the quality management system, in order to enable the notified body to assess the technical documentation, the provider must submit an application to the notified body. This application must contain the provider's identification data, the technical documentation and a declaration that the provider has not lodged this application with another notified body. In the case of SMEs, notified bodies must provide them, on request, with the simplified technical documentation form developed by the European Commission[24].

Upon receipt of the application, the notified body is responsible for assessing the technical documentation. The AIA envisages different situations where the notified body is empowered to carry out other actions when it considers that the documentation provided is not sufficient. In this sense, when the body considers it necessary, it may: access the set of training, validation, and test data used, access the training model and the trained model of the AI system, oblige the provider to carry out additional tests or, if necessary, carry them out itself. We understand that in all these cases the notified body must justify the reasons why it considers it necessary to carry out these actions that go beyond the access to the technical documentation initially provided by the provider.

This set of activities will help the notified body to adequately verify the compliance of an AI system with the requirements of the AIA.

Upon completion of the assessment of the technical documentation by the notified body according to the processes described above, the notified body shall notify the provider or, where applicable, the authorised representative of the decision taken. This decision shall indicate whether or not the assessed AI system complies with the requirements of the AIA. In cases where the decision is positive, the notified body will issue the EU certificate for the technical documentation. If the notified body does not consider that the AI system complies with the Act, the notified body shall indicate this. Where the refusal has occurred because the data used is considered not to comply with the requirements of the Act, the notified body shall set out specific considerations on the data used during the training of the AI system and oblige the provider to undertake further training before the provider reapplies for a new assessment of its AI system[25].

Once all the above mentioned actions have been carried out and a favourable decision has been obtained for all of them, the conformity assessment shall be deemed to have been passed.

[24] Article 11.1 European Regulation on Artificial Intelligence.

[25] See Section 4.6 Annex VII of the AI Regulation.

V. Conformity assessment according to the type of Artificial Intelligence system

Article 43 of the AIA sets out the conformity assessment process to be carried out by each provider taking into account the type of AI system it has designed or is designing.

The AIA establishes two broad groups of high-risk AI systems. On the one hand, those AI systems that are used for a number of purposes (high risk purposes, Annex III), and on the other hand, those AI systems that are products or safety components of products that are subject to harmonisation legislation and that such legislation provides for the conformity assessment of these products to be carried out by a notified body (high risk products, Annex I)[26].

Depending on the type of AI system, one conformity assessment process or another will be applied.

1. Artificial Intelligence Systems whose purpose is considered high-risk (high-risk purposes)

Annex III of the European Artificial Intelligence Regulation establishes a list of purposes that are considered high-risk when carried out by an AI system. The conformity assessment process differs in part according to the type of purpose for which the AI system is intended to be used. Thus, we have to distinguish the conformity assessment that is envisaged on the one hand for high-risk systems whose purpose is biometric identification and, on the other hand, the rest of the purposes envisaged in this Annex.

It is now time to analyse these differences, which are shown in the following table.

[26] See Article 6 together with Annexes I and III of the AI Regulation.

Forms of conformity assessment.	High-risk purposes (Annex III)
Self-assessment (Annex VI) or presence of notified body (Annex VII)	Biometric identification.
Self-assessment (Annex VI)	Infrastructure. Education and vocational training. Employment, employee management and access to self-employment. Access to and enjoyment of essential private services and essential public services and benefits. Law enforcement by police authorities. Migration, asylum and border control management. Administration of Justice and democratic processes.

1.1. Artificial Intelligence Systems whose high-risk purpose is biometric identification

According to Article 43.1 of the AIA, the conformity assessment process for biometric identification systems may be either self-assessment or the presence of the notified body. This will depend on whether or not the provider has implemented harmonised standards or common specifications in his AI system to meet the requirements of the AIA.

Therefore, where the provider has applied harmonised standards or common specifications, the provider has the choice between carrying out the self-assessment of conformity (Annex VI) or requesting a notified body to carry out the self-assessment of conformity (Annex VII).

However, if the provider does not use harmonised standards or applies them only partially or does not have common specifications to comply with the requirements of the Act, the provider must necessarily go through the conformity assessment process with the presence of a notified body.

1.2. Artificial Intelligence Systems whose high-risk purpose is other than biometric identification

For all other purposes considered high-risk other than biometric identification, providers shall carry out the self-assessment of compliance as set out in Annex VI as explained above[27].

Although in these cases the verification process is carried out entirely by the providers, the verification process must be fully documented and always available to the market surveillance authority. In this regard, providers are

[27] Article 43(2) of the European Artificial Intelligence Regulation.

obliged to demonstrate, upon reasoned request of the competent authority, the conformity of their AI system with the essential requirements of the AIA[28].

2 Artificial Intelligence systems for products or safety components of products considered to be of high risk

Annex I refers to a number of products and product safety components whose design, placing on the market, and surveillance are regulated by harmonisation laws that follow the NLF. As noted above, each of these laws has a similar structure. Within these common elements, all these products have to pass the conformity assessment that is provided for in these legal texts. Each legislation provides for various conformity assessment processes depending on the characteristics of each of these products.

Where an AI system is a product in itself or a safety component of one of these products, the conformity assessment of these AI systems shall consist of two main processes.

On the one hand, the conformity assessment process of this AI system will be the one envisaged for that product or product safety component in the applicable legislation. On the other hand, this conformity assessment process must incorporate some of the actions we have already mentioned related to the technical documentation to be reviewed by the notified bodies and which derive directly from the AIA itself.

2.1. Conformity assessment of Artificial Intelligence systems in the structure of harmonisation legislation

The verification of AIA requirements will be part of the conformity assessment process foreseen in each of the harmonisation laws for those products or product safety components. The aim is that if an organisation intends to develop a toy, a medical device or a lift with an integrated AI system, it will not have to carry out two conformity assessments, but only one, i.e., the one indicated in the harmonisation legislation applicable to those products[29].

The verification process foreseen in each act of harmonisation legislation depends in most cases on the product. In other words, different conformity assessment processes are foreseen within the same piece of legislation[30].

[28] Article 16(k) of the European Artificial Intelligence Regulation.

[29] Recital 124 of the AI Regulation.

[30] Annex II. Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products and repealing Council Decision 93/465/EEC.

Conformity assessment processes under the different EU harmonisation acts	
Module A	Internal control of production (self-assessment) Internal production control plus supervised product testing Internal manufacturing control plus supervised product control at random intervals
Module B	EC type-examination
Module C	Conformity to type based on internal production control
Module C1	Conformity to type based on internal production control plus supervised product testing
Module C2	Conformity to type based on internal production control plus supervised control of products at random intervals.
Module D	Conformity to type based on quality assurance of the production process
Module D1	Quality assurance of the production process
Module E	Conformity to type based on product quality assurance
Module E1	Quality assurance of inspection and testing of the finished product
Module F	Conformity to type based on product verification
Module F1	Conformity based on product verification
Module G	Compliance based on unit verification
Module H	Conformity based on full quality assurance
Module H1	Conformity based on full quality assurance plus design review

The way to integrate the verification of the requirements foreseen in the AIA into the conformity assessment methodology designed by each harmonisation legislation for the different products is not made explicit in this standard throughout its articles.

However, Recital 64 calls for a simultaneous and complementary application of the various pieces of legislation that may be applicable in these cases. The aim is none other than to avoid unnecessary burdens or costs. The verification of these requirements should follow the same philosophy.

It is very likely that the integration of these requirements will eventually be deployed through the development of harmonised standards or common specifications. It is also possible that the various harmonisation laws, as they are amended or updated, will begin to provide for the integration of AIA requirements into these laws. In this respect, the Machinery Regulation 2023/1230 has already partly achieved this process of understanding between the different regulatory texts[31].

[31] Recital 54, Article 25.2 and Annex I. Part A of Regulation (EU) 2023/1230 of the Europe-

In the absence of the drafting of such harmonised standards or other instruments that would partially assist or facilitate the integration of AIA requirements into the assessment process of products or safety components of products that already provide for a conformity assessment process, the conclusion to be drawn is that the integration of AIA requirements cannot affect the logic, methodology, or structure indicated in such harmonisation laws.

2.2. Verification obligations arising from the Act on conformity assessment of products or product safety components

While the conformity assessment process to be followed will be marked by harmonization legislation, in order to ensure that the AI system complies with the requirements of the AIA, it enables notified bodies in accordance with the harmonisation legislation of the product in question to carry out a number of actions that follow from the conformity assessment process foreseen in the AIA.

In particular, the notified body intending to carry out the assessment of an AI system, which is itself a product or a safety component, shall carry out the following actions: it shall have access to the technical documentation of the AI system, it shall have access to the training, validation, and test data set used, and it shall have access to the training model and the trained model of the AI system. In addition, the notified body shall refuse to issue an EU certificate of assessment of the technical documentation if it finds that the system does not meet the requirements concerning the data used for training[32].

The latter is to ensure that notified bodies can carry out an appropriate conformity assessment procedure which, while respecting the structure of the harmonisation legislation in question, is adequate and in line with the requirements of the AIA.

Forms of conformity assessment	**Conformity assessment according to type of AI system** (Article 43)
Self-assessment	High-risk purposes. Annex III. Paragraphs 2 to 8 (all except biometric identification).
Self-assessment or Presence of notified body	High-risk purposes. Annex III. Paragraph 1. (biometric identification only)
Conformity assessment according to product harmonisation legislation	High Risk Products. Annex I

an Parliament and of the Council of 14 June 2023 on machinery, and repealing Directive 2006/42/EC of the European Parliament and of the Council and Council Directive 73/361/EEC.

[32] Article 43.3.

2.3. Possibility of dispensing with the presence of the notified body

Where an AI system is a product or a safety component of a product under the harmonisation legislation, the manufacturer may dispense with the conformity assessment in the presence of a notified body as long as two cumulative conditions are met.

Firstly, the act of harmonisation legislation should provide for the possibility for the manufacturer to dispense with the third party verification process if the manufacturer has applied harmonised standards covering the requirements of that legislation.

Secondly, the manufacturer has applied harmonised standards or common specifications covering the requirements for high-risk AI systems.

Ultimately, the presence of a notified body may be dispensed with where this is provided for in the legislative act to which the product is subject and where harmonised standards or common specifications covering all the requirements of both the legislative act to which the product is subject and the requirements of the AIA-compliant AI system have been applied.

There are currently no harmonised standards or common specifications covering the requirements set out in the AIA for AI systems, so this option of circumventing the indicated third party conformity assessment is not yet possible.

For these cases, we understand that, even if the manufacturer could avoid the conformity assessment of his product by a Notified Body under harmonisation legislation by applying harmonised standards covering the requirements of that standard, the notified body would have to intervene if there are no harmonised standards or common specifications covering the requirements of the AIA.

For example, Article 19.2 of Directive 2009/48/EC on the safety of toys provides for the possibility for the manufacturer to dispense with conformity assessment by notified bodies if the manufacturer has applied harmonised standards covering all the relevant requirements of that Directive.

Therefore, if an AI system is a toy or a safety component of a toy, given that there are currently no harmonised standards or common specifications covering the requirements set out in the AIA, the manufacturer of that toy will not be able to circumvent the notified body even if the toy harmonisation legislation allows it.

AI Product/System	Have you applied harmonised standards covered by the legislation?	Notified body required?
Toy	Does not cover the requirements of the Toys Directive	Yes
Toy	It does cover the requirements of the Toys Directive	No
Toy that is also an AI system	It does cover the requirements of the Toys Directive	Yes
	Does not cover AIA requirements	
Toy that is also an AI system	It does cover the requirements of the Toys Directive	No
	It does meet the requirements of the AIA	

VI. Notified bodies responsible for carrying out the conformity assessment of the different Artificial Intelligence systems.

As a general rule, providers of AI systems that are to undergo a conformity assessment process in the presence of a notified body may choose the one they deem appropriate, provided, of course, that it has been accredited to be able to carry out conformity assessment of AI systems.

However, there are some specific rules where providers are limited in their freedom to choose one notified body over another.

Firstly, according to the last paragraph of Article 43.1 of the AIA, an AI system intended for biometric identification and intended for use by law enforcement or immigration authorities, the notified body shall be[33]: a) either the competent data supervisory authorities under Regulation (EU) 2016/679 or Directive (EU) 2016/680 (currently the AEPD)[34], b) or any other authority designated pursuant to the same conditions laid down in Articles 41 to 44 of Directive (EU) 2016/680[35]. We therefore understand that the notified body

[33] Article 74.8 of the Artificial Intelligence Regulation.

[34] This applies both to police activities and, where appropAIAte, to immigration and border control activities.

[35] These articles refer to the need for such a public authority in the framework of its activities to enjoy a high degree of independence. Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data and repealing Council Framework Decision 2008/977/JHA.

for these AI systems will be the AEPD (Spanish Data Protection Agency) or an analogous body in data protection matters, as is the case with the GCJ in matters of justice[36], as it is very unlikely that an authority with the same degree of independence and the same powers of supervision and access to data as the current AEPD will be created in Spain in these contexts.

We do not believe that the Spanish Agency for the Supervision of Artificial Intelligence (AESIA) currently has the level of independence required by Articles 41 to 44 of Directive 2016/680 and advocated by the AIA for these cases. In this regard, unlike the AEPD, which has been configured as an independent administrative authority in our domestic law[37], the AESIA is constituted as a state agency with autonomy and technical independence[38], however, from its statutes it can be glimpsed that the central government has a lot of control capacity, especially when it comes to choosing the essential bodies of this agency[39].

Secondly, also according to the last paragraph of Article 43.1 of the AIA, providers whose AI system is intended for biometric identification and such a system is intended to be put into service by EU institutions or agencies, the notified body shall be the European Data Protection Supervisor[40].

Thirdly, Article 41.3 establishes that the notified bodies that are competent to carry out conformity assessment of products subject to the harmonisation legislative acts shall have the power to verify conformity with the requirements of the AIA when the product is an AI system. These notified bodies will not only have to comply with the requirements foreseen in the harmonisation legislative acts that enable them to carry out conformity assessments, but they will also have to implement another series of requirements that derive from the European AI Regulation in order to be assessors of the requirements foreseen in this standard. These requirements include the need for these bodies to have sufficiently competent personnel and resources to be able to adequately verify the requirements of the AIA[41].

A priori, it is logical to think that the notified body that carries out the

[36] The data protection authority in the sphere of the Administration of Justice is the General Council of the Judiciary. Article 236. Nonies. Organic Law 6/1985, of 1 July 1985, on the Judiciary.

[37] Article 109 of Law 40/2015, of 1 October, on the Legal Regime of the Public Sector.

[38] Article 108 bis of Law 40/2015, of 1 October, on the Legal Regime of the Public Sector.

[39] Such as the president of AESIA or the management of AESIA. See Royal Decree 729/2023 of 22 August, approving the Statute of the Spanish Agency for the Supervision of Artificial Intelligence.

[40] Article 74.9 of the AIA.

[41] Article 31(4), (9) and (10) of the AI Regulation.

conformity assessments of a lift or a toy in accordance with its regulations is the most qualified to carry out the conformity assessments of the same lift or toy when it has an integrated AI system as a safety component. This is because these notified bodies have extensive experience in verifying the requirements for lifts, however, it will be necessary for these bodies to put in place the human and technical means to be able to effectively verify the requirements deriving from the AIA as well.

The AIA enables these bodies to outsource the verification of AIA requirements to third parties[42]. It is therefore likely that notified bodies of products or product safety components subject to harmonisation legislation that intend to extend their verification processes to AI systems that are themselves products or product safety components for which they have been carrying out conformity assessments will eventually subcontract these activities to entities specialised in the verification of AIA requirements.

AI system	Notified Body
Biometric identification used by immigration or law enforcement authorities	Quite possibly the AEPD
Biometric identification used by European authorities and institutions	European Data Protection Supervisor
Product or product safety component subject to harmonisation legislation	Notified body of the product or product safety component.

VII. Cases where conformity assessment is not required or presumptions of conformity of compliance exist.

The AIA provides for a number of situations where an AI system does not need to undergo a verification of conformity process prior to being placed on the market. These cases are exceptional and justified. At the same time, it also regulates some cases where the AI system is presumed to be in conformity with the requirements of the AIA.

1. Prior authorisation to place Artificial Intelligence systems on the market

According to Article 46.1 of the AIA, any market surveillance authority may authorise the placing on the market of an AI system without such a system having undergone a conformity assessment process.

[42] Article 33 of the AI Regulation.

This situation will be allowed when the use of this AI system is necessary to protect human life or health, the environment, public safety, or critical industry and infrastructure assets. As can be seen, the reasons for which such authorisations may be granted cover a wide range of scenarios. However, the AIA contains certain provisions that seek to guarantee the exceptional nature of this type of measure.

First, the market surveillance authority must ensure that the AI system complies with the minimum essential requirements for AI systems[43] before granting authorisation. The authorisation must be duly reasoned, taking into account the reasons for which the authorisation has been granted.

Secondly, once the authorisation has been granted, the market surveillance authority will communicate it to the European Commission and the other Member States. Within 15 days, the latter may raise any objection to the authorisation granted on the grounds that it is not sufficiently justified or is contrary to European Union law. If no objections are raised, the authorisation will be deemed to be justified; otherwise, the Commission will initiate the appropriate procedures to communicate with the supervisory authority that granted the authorisation and the system operators so that they can express their opinion. The Commission will finally decide whether or not the authorisation is justified.

The involvement of the Commission or the Member States in this process is due to the fact that any surveillance authority in any Member State can authorise an AI system to be used anywhere in the European Union. We therefore believe that this is a control measure on the part of the Member States and the European Commission towards the various market surveillance authorities that may grant this type of exceptional authorisations with a certain degree of ease or laxity.

2. Placing the Artificial Intelligence system on the market without prior authorisation

In duly justified emergencies on exceptional grounds of public security or in the event of a specific, substantial, and imminent threat to the life or physical safety of natural persons, a high-risk AI system may be brought into operation by the public order authorities or civil protection authorities without the need to obtain the previously mentioned authorisation.

[43] Article 46.3 states that authorisation will only be issued "*if the market surveillance authority concludes that the high-risk AI system complies with the requirements of Section 2*". This is to be understood as referring to Section 2 of Chapter III of the AI Regulation, the section covering the essential requirements, i.e. Articles 8 to 15.

In such cases, the authorities using these systems must necessarily apply without undue delay to the market surveillance authority for authorisation as explained in the previous paragraph. If the market surveillance authority refuses authorisation, the use of the system must be suspended and the results and output information derived from such use must be discarded.

3. Exemptions from conformity assessment for products subject to harmonisation legislation

Where an AI system is a product or a safety component of a product subject to harmonisation legislation, procedures for exemption from conformity assessment shall only be allowed if provided for in the applicable harmonisation legislation.

4. Presumptions of Conformity

The AIA establishes a number of presumptions of conformity that refer to various requirements for AI systems. Thus, if a provider applies such a requirement in accordance with that presumption, it must be deemed to comply with the AIA requirement.

There are two presumptions in Article 42: on the one hand, when AI systems have been trained and tested with data reflecting the specific geographical, behavioural, or functional environment of their use, it shall be presumed that the system complies with the requirements of Article 10.4 of the AIA. We understand that the presumption of conformity provided for in this provision has the same effects as the presumption of conformity of harmonised standards or common specifications, i.e., the use of these does not circumvent the conformity assessment of the requirements covered by this presumption, but it does ensure a much faster verification process than in cases where such requirements would have to be demonstrated by the application of other techniques or standards.

On the other hand, where an AI system has been certified or issued with a declaration of conformity with a cybersecurity scheme and whose references have been published in the OJEU[44], compliance with the requirements of Article 15 of the AIA on cybersecurity shall be presumed to the extent that such

[44] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (European Union Agency for Cybersecurity) and the certification of information and communication technology cybersecurity and repealing Regulation (EU) No 526/2013 (Cybersecurity Regulation).

requirements are provided for in that cybersecurity certificate or declaration of conformity.

Finally, although not a presumption of conformity in the strict sense, Article 57.7 of the AIA provides for the possibility for market surveillance authorities or notified bodies to take into account positively, for the purpose of accelerating the conformity assessment procedure of an AI system, reports that have been issued on that system due to its participation in a controlled test area.

VIII. Reflections on the conformity assessment regulation set out in the AI Act

As has been pointed out throughout this chapter, the conformity assessment process is part of a battery of measures and instruments that aim to ensure that a product placed on the market complies with minimum safety and guarantee requirements for people.

The process of verifying the requirements for AI systems is therefore shown to be an elementary step in ensuring that such products are suitable for use.

The European legislator has considered that for a significant number of high-risk AI systems, this verification of the requirements should be carried out by the provider itself[45]. Workplace, education, a large part of the public sector, and police and judicial proceedings, except for biometric identification, are covered by the internal control of the providers.

This legislative decision is surprising to say the least. Despite the efforts to design a legal framework that is clearly innovative and respectful of the fundamental rights present during the lifecycle of AI systems through the obligation to impose requirements that reduce the main risks associated with such systems, self-assessment by providers may generate significant suspicions as to the effective integration of these requirements in AI systems operating in the European Union.

The European legislator is therefore relying on the industry in these early years to self-verify, if necessary, while awaiting the development of increasingly mature products that meet legislative requirements.

It is true, however, that despite self-assessment as a general rule, the AIA places great weight on Member States and the European Commission to ensure effective compliance with this new standard.

[45] See Annex III except paragraph 1 regarding biometric identification.

Thus, firstly, the more or less leading role given by Member States to market surveillance authorities will be essential. These public authorities should have sufficient human and technical means to carry out an effective supervision of the AI systems placed on the market, whether they have undergone a self-assessment process or whether notified bodies have been involved. The independence of these authorities from the respective governments and private sector entities should be adequate, not forgetting that these authorities will oversee both public and private sector systems.

Secondly, the AIA itself gives the European Commission the possibility, through delegated acts, to change the self-assessment currently required for high-risk AI systems in Annex III to assessment by a notified body[46].

IX. Conclusions

1. Any AI system considered to be of high-risk that intends to be placed on the EU market must undergo a prior verification process to ensure compliance and integration of the essential requirements demanded by the AIA.

2. The AIA provides for two conformity assessment processes: on the one hand, self-assessment, which will be carried out by the provider of the AI system itself, and, on the other hand, assessment in the presence of a notified body, i.e., a third party entity that has been accredited to carry out conformity assessments of AI systems in accordance with this European regulation.

3. The choice of one conformity assessment process or another is not left to the provider but will depend on the type of AI system that has been developed.

4. Where the AI system is intended for biometric identification, the provider may choose one or the other conformity assessment process provided that he has applied harmonised standards or common specifications to implement the essential requirements of the AIA. Otherwise, the provider shall use the conformity assessment process in the presence of a notified body.

5. Where the AI system is intended for some of the purposes considered as high-risk by the AIA(except biometric identification), the conformity assessment of the AI system (the self-assessment) shall be carried out by the provider itself.

6. When the AI system is a product or a safety component of a product that is covered by European legislation, the conformity assessment process shall be that provided for in the standard governing the product. The incor-

[46] Article 43.6 of the European Artificial Intelligence Regulation.

poration of the essential requirements of the AIA shall not alter the structure of the assessment process provided for in the standard governing the product, although the requirements of the AIA shall be verified in that process.

7. As a general rule, providers are free to choose the notified body that will assess their AI systems, however, on certain occasions where the public sector is involved, they will not have a choice.

8. As a general rule, any AI system intended to be placed on the EU market must have undergone a conformity assessment process, however, there are exceptional cases where such an initial verification process is not required

GENERAL REGIME OF OBLIGATIONS FOR PROVIDERS AND DEPLOYERS IN THE ARTIFICIAL INTELLIGENCE ACT

Adrián Palma Ortigosa

Lecturer in the Department of Administrative Law of the Universitat de València[1]

I. Introduction

This paper studies different precepts of the AIA. Firstly, it looks in depth at the different operators involved along the AI value chain, defining each of these agents and analysing their main functions and obligations. Secondly, the role and functions assigned to one of the competent authorities for AIA compliance, i.e., the notifying authority, are examined. Thirdly, it analyses the measures envisaged in this European standard in favour of SMEs and start-ups with the aim of facilitating the correct adaptation of the latter to this standard. Fourthly, it examines the procedure for notifying serious incidents of AI systems regulated in the AIA[2].

[1] This work has been carried out in the framework of the following research projects:

"Algorithmical Law" (PROMETEO/2021/009. Funded by the Generalitat Valenciana. "Regulation of the digital economy: public protection of equality and algorithmic tools" (PID2019-108745GB-I00). Ministry of Science and Innovation. "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by the Ministry of Science and Innovation. "Algorithmic Tools for Citizens and Public Administrations" (Proyectos de Generación de Conocimiento, Ministerio de ciencia e Innovación, convocatoria 2021, PID2021-126881OB-I00). "Algorithmic Decisions and the Law: Opening the Black Box" (TED2021-131472A-I00) of the Recovery, Transformation and Resilience Plan. Digital Rights Agreement-SEDIA Scope 5 and 6 (2023/C046/00228673).

[2] This work has been carried out in the framework of the following research projects: "Algorithmical Law" (PROMETEO/2021/009. Funded by the Generalitat Valenciana. "Regulation of the digital economy: public guardianship of equality and algorithmic tools" (PID2019-108745GB-I00). Ministry of Science and Innovation. "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by the Ministry of Science and Innovation. "Algorithmic tools for citizens and Public Administrations" (Proyectos de Generación de Conocimiento, Ministerio de ciencia e Innovación, convocatoria 2021, PID2021-126881OB-I00).

II. Development, processing, and final content of the articles of the Act involved

Some of the precepts analysed in this part of the collective work have undergone important changes throughout the AIA legislative process. Among the main changes we can highlight the following.

Firstly, in terms of content, the obligations imposed on the various operators involved in the value chain of AI systems have been considerably increased. Particularly noteworthy is the increase in the requirements for the deployer. This actor was initially referred to as the user, but in the latest known versions of the Act this name has been replaced. In our opinion, this change is the right one, as the term user was often confused with the persons affected by the decisions. With regard to the obligations, it is worth highlighting the obligation to carry out a fundamental rights impact assessment. This requirement was implemented in the last agreements of the Act, but its origin stems from the amendments made by the European Parliament[3].

Secondly, in terms of form, some of these precepts have been moved or merged, although the place where they were located in the first version of the AIA has not been greatly affected.

III. Operators in the value chain of Artificial Intelligence systems in the Act

Throughout the complex value chain of AI systems, all kinds of actors may be present and involved in one way or another. The AIA has tried to give names and surnames to all these actors, which it calls operators. In terms of names, it provides a definition of each of them; in terms of surnames, it assigns each of these operators a set of roles and obligations.

The objective of the Act is twofold: on the one hand, it assigns clear responsibilities, so that each actor knows what he has or does not have to do under the Act, and on the other hand, it reduces or mitigates the possibility of evading potential liabilities when the system commits any damage to third parties.

It is now time to analyse each of these operators and their functions[4].

[3] See amendment 413 of the European Parliament resolution of 14 June 2023 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union.

[4] These are: provider, product manufacturer, deployer, authorised representative, importer or distributor. Article 3.8. AI Regulation.

1. The provider and his obligations

The provider is any natural person, legal entity, or public authority that develops an AI system or AI model for general use or for which an AI system or AI model for general use is developed and places it on the market or puts the AI system into service under its own name or trademark[5].

The provider is the central axis on which the greatest number of obligations and responsibilities rest in the compliance with the AIA. This is due to the fact that it is the subject in charge of developing or mandating the development of the AI system, which will subsequently be used in decision-making.

These obligations are mentioned throughout the regulatory text, although they are concentrated in the initial precepts of the regulation.

Main obligations	**Article**
Essential requirements for AI systems	8 a 15
Essential obligations: technical documentation, conformity assessment, CE marking, etc.	16
Quality management system	17
Preservation of documentation	18
Log files	19
Remedial action and reporting obligations	20
Cooperation with the competent authorities	21

In this work we will only look at the obligations mentioned in Articles 20 and 21, i.e., corrective measures, reporting obligations, and cooperation with competent authorities. The rest of the obligations are dealt with in other sections of this collective work.

Firstly, as regards the implementation of corrective measures, a provider who considers or has reason to consider that an AI system he has introduced is not in conformity with the AIA, shall take appropriate measures to deactivate or recall it. In addition, it shall inform the distributors of the system and, where applicable, those responsible for the deployment of the system, as well as authorised representatives and importers, of this situation.

Secondly, if the provider considers that an AI system presents a risk following a notification from the person responsible for the deployment of such a system, the provider shall investigate the causes of the risk and inform the

[5] Article 3.3. AI Regulation.

competent market surveillance authorities and, where appropriate, the notified body that issued the relevant conformity assessment certificate.

Thirdly, providers are obliged to provide all information and documentation requested by the competent authority to demonstrate compliance with the AIA. Provided that there is a reasoned request from the competent authority, the provider will also give it access to the automatically generated records archive of the system to the extent that these records are under its control. These duties of information and cooperation are essential, as in many cases, AI systems that are introduced into the market do not go through any kind of control beyond the system provider's own internal self-assessment, hence the role of market surveillance authorities once the system is making decisions is very relevant.

2. The deployer and his obligations

The deployer is any natural or legal person or public authority using an AI system under its own authority, except when the use is in the context of a personal activity of a non-professional nature[6].

Along with the provider, the deployer is the operator on whom the AIA imposes the most compliance obligations. This makes sense given that it will be the one using the system for much of the time it is in operation.

The obligations of the person responsible for deployment are mainly concentrated in Articles 26 and 27 of the AIA. In order to facilitate the reading of these precepts in this paper, we have grouped the content of these obligations into four main groups.

Obligation	**Article and paragraph**
Compliance with essential requirements	Article 26(1), (2), (3), (4) and (6).
Other compliance obligations	Article 26(5), (7), (8), (9), (11) and (12)
Obligations related to the use of biometric identification systems	Article 26 paragraph 10.
The Fundamental Rights Impact Assessment	Article 27

2.1. Obligations related to compliance with the essential requirements of the Act

The AIA sets out a number of essential minimum requirements to be integrated by the provider during the development of AI systems. These minimum requirements are recognised in Articles 8 to 15 of the AIA. Part of the

[6] Article 3.4. AI Regulation.

content of these requirements is designed to enable the deployer to make proper use of the AI system that the provider has designed.

First, the deployer must take appropriate technical and organisational measures to ensure that they use the system in accordance with the system's instructions for use. Instructions for use are a common tool provided for in European product standards that help to reduce opacity and promote transparency in the operation of such products[7], in our case AI systems[8]. If those responsible for deployment do not properly follow these instructions, they are likely to be held liable for damage caused by the AI system.

Secondly, it is stipulated that deployers shall entrust the tasks of monitoring the AI systems to persons with the necessary competence[9], training, and authority. Although the provider must have designed the system in such a way as to facilitate effective oversight of the system[10], it is up to the user to designate competent and trained personnel for this purpose. These human oversight measures shall be implemented without prejudice to any other measures that the deployer may be required to implement under other rules of national or European law[11]. For example, Article 22 of the General Data Protection Regulation obliges entities using AI systems in fully automated decision-making to implement human oversight processes after the decision has been taken by the algorithm[12]. To the extent possible, and where compatible, the two obligations could complement each other.

Third, the deployer must ensure that the input data used by the system are relevant and sufficiently representative for the intended purpose of the system. It is clarified that this obligation will come into play to the extent that the deployer exercises control over such data. It should be noted that the same obligation applies to the providers of AI systems[13], however, as soon as the deployer takes over the use of the AI system and is the one who inputs

[7] Instruction for use is to be understood as "*information provided by the provider to inform the deployer, in particular, of the intended purpose and proper use of an AI system*". Article 3.15 of the AI Regulation.

[8] The minimum content of the instructions for use is defined in Article 13.3 of the AI Regulation.

[9] On human supervision see among others: Lazcoz Moratinos, G and Obregón Fernández, A., "La supervisión humana de los sistemas de inteligencia artificial de alto riesgo. Aportaciones desde el Derecho Internacional Humanitario y el Derecho de la Unión Europea". *Revista electrónica de estudios internacionales*, n.º 42, 2021.

[10] See Article 14 of the AI Regulation (Human surveillance).

[11] Article 26.3 of the AI Regulation.

[12] Palma Ortigosa, A., *Automated decisions and data protection. Special attention to Artificial Intelligence systems*. Dykinson. Madrid. 2022, p.286 and ff.

[13] Article 10(3) and (4) of the AI Regulation (Data).

the data, it should be understood that the obligations in this area may change, in particular if the deployer has control over these data and no longer uses data that have the specific characteristics indicated for the purpose of the AI system.

Fourthly, the deployer should retain the log files automatically generated by the system for at least 6 months after they are generated, provided that the log files are under their control. It is up to the provider to design the system so that it automatically generates such records[14], the key element will be to check who has control over them.

2.2. Other obligations arising from compliance with the Act

In addition to the specific obligations relating to the essential requirements for AI systems, the AIA places a number of obligations on users that seek to ensure adequate compliance of that AI system with this European standard.

First, there is an obligation to monitor the functioning of the AI system on the basis of the instructions for use and, where appropriate, to inform the provider about the post-market surveillance system[15]. In addition, if the deployer identifies that the system could present a risk, he must inform the provider or distributor and the relevant market surveillance authority[16]. In turn, if the deployer detects a serious incident in the AI system, the deployer shall report the incident to the provider, then to the importer or distributor and to the relevant market surveillance authority. If the deployer is unable to contact the provider, the deployer should take all the communications and actions required of the provider in the event of a serious incident as set out in the AIA[17] .

Secondly, the person responsible for deployment shall, before using the AI system in the workplace, inform the workers' representatives and workers exposed to the use of the AI system of its implementation. This information shall be provided, where appropriate, in accordance with the channels provided for in national or European law governing information processes for the benefit of workers and their representatives. In this respect, there are already some rules which oblige employers to inform work-

[14] Article 12. AI Regulation (Registers).

[15] The post-marketing surveillance system is regulated in Article 72 of the AI Regulation.

[16] For more information on the risk that an AI system may present, see Article 79 of the AI Regulation.

[17] For more information on the latter, see Section VI of this paper and Article 73 of the AI Regulation.

ers' representatives[18] and in some cases also workers themselves about the use of algorithmic systems and the consequences of their use[19]. As far as possible, these information obligations should be integrated and complemented.

Thirdly, it is up to the deployers who are public authorities to register AI systems in the EU database created by the AIA[20]. In the event that such a system is not registered in this database, they must inform the provider or distributor of this fact.

Fourth, deployers will use the information given by the provider on the AI system to carry out the impact assessment required by EU data protection law[21]. Of course, this obligation will come into play when the deployer will use personal data. In this sense, much of the information that has been documented during the design process of the AI system by the provider will be essential to comply not only with this specific obligation, but also with other obligations under data protection law, such as the principle of privacy by design.

Fifth, those responsible for the deployment of AI systems whose purpose is considered to be of high risk[22] must inform natural persons that such systems are making fully or partially automated decisions about those persons. Unlike the General Data Protection Regulation, which provides for a more protective regime for fully automated decisions as opposed to partially automated decisions[23], the AIA focuses primarily on the type of AI system being used and not on the more or less active involvement of the individual in the decision-making process of the AI system.

In the case of AI systems whose purpose is law enforcement, the information to be provided will be that indicated in Article 13 of the Police Per-

[18] This is contemplated in Article 64.4.d). Royal Legislative Decree 2/2015, of 23 October, approving the revised text of the Workers' Statute Law.

[19] Proposal for a Directive of the European Parliament and of the Council on improving working conditions at work on digital platforms.

[20] For more information on this register see Article 49 of the AI Regulation. For more information on the database see Article 71.

[21] See Article 35 of Regulation (EU) 2016/679 or Article 27 of Directive (EU) 2016/680.

[22] Article 6.2 and Annex III of the AI Regulation.

[23] See Articles 13.2.f), 14.2.g), 15.1.h) and 22. For an analysis of these precepts, see: Cotino Hueso, L., "Derechos y garantías ante el uso público y privado de inteligencia artificial, robótica y big data", in Bauzá, Marcelo (dir.), *El Derecho de las TIC en Iberoamérica, Obra Colectiva de FIADI*. La Ley- Thompson-Reuters, Montevideo, pp. 917-952.

sonal Data Directive[24], a European text which in Spain has been transposed by Organic Law 7/2021[25].

Sixth, a general duty to cooperate with the competent national authorities in relation to compliance with the AIA is established.

2.3. Specific obligations when using an Artificial Intelligence system for biometric identification purposes

The AIA provides for a number of actions to be carried out by the deployer when intending to use a high-risk AI system of delayed targeted biometric identification.

These actions seek to ensure that the use of these AI systems, which conduct significant risks of use, is carried out with a minimum of safeguards. These guarantees include: the need to request prior authorisation to use these systems for this purpose, the non-indiscriminate use of the AI system on individuals, the need to submit annual reports to market surveillance authorities, as well as data protection authorities on the use made of these tools[26].

2.4. The obligation to carry out a fundamental rights impact assessment

As initially indicated, this obligation was not contained in the first versions of the original text of the AIA. It was the European Parliament that initially opted to introduce this measure as a general guarantee in favour of those affected by whom AI systems will make decisions[27]. Although with some changes, the final wording of this provision is very similar to that proposed by the European Parliament, although the level of requirements has been reduced[28].

The obligation to carry out an impact assessment rests with certain deployers, namely:

[24] Article 13 Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data and repealing Council Framework Decision 2008/977/JHA.

[25] Article 21 of Organic Law 7/2021 of 26 May on the protection of personal data processed for the purposes of the prevention, detection, investigation and prosecution of criminal offences and the execution of criminal penalties.

[26] Article 26.10 of the AI Regulation.

[27] See amendment 413 of the European Parliament resolution of 14 June 2023 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union.

[28] An approach to different models of impact assessments of AI systems can be found in: Simón Castellano, P., *La evaluación de impacto algorítmico en los derechos fundamentales,* Aranzadi, 2023.

On the one hand, any deployer that is a public authority or private entity providing services to such authorities where the high-risk AI system is used for one of the purposes listed in Annex III of the Regulation. Only high-risk AI systems that are used as a security component of the management and operation of critical digital infrastructures are excluded[29].

On the other hand, any deployer, regardless of whether or not it is a public authority, who uses its AI system to assess the creditworthiness of natural persons or to establish their credit rating or uses such a system for risk assessment and pricing in relation to natural persons in the case of life and health insurance.

In both cases, only the deployers implementing the first use of the AI system are required to carry out this impact assessment. In subsequent uses, the deployer may rely on the initial impact assessment that was carried out unless any of the elements to be present in the assessment have been modified or altered by the use of the system. In the latter case, the deployer shall update the impact assessment to the extent that it has been modified by changes or alterations to the AI system.

The impact assessment will consist of a series of actions that are closely linked to any information provided by the provider to the AI system deployer.

First, a description of the processes that the deployer will carry out in which he will operate the AI system, as well as the length of time for which he intends to use it and its frequency of use, should be provided.

Secondly, the categories of natural persons and groups that may be affected, as well as the specific risks that may affect them during the use of the AI system, will be established. This information may, of course, be partly compiled from documentation provided by the provider under the transparency obligations imposed on the latter by the AIA[30].

Thirdly, a description of any human oversight measures intended to be deployed to use the system, as well as any other measures aimed at reducing the potential risks that the use of the system may generate, must be provided. On this last point, the AIA explicitly obliges those responsible for deployment to establish internal governance arrangements and complaints mechanisms[31].

To facilitate the preparation of the impact assessment, the AI Office will develop a simplified questionnaire. In addition, if the controller has already carried out an impact assessment under personal data protection law, the im-

[29] Critical infrastructures such as road traffic, water, gas, heating, electricity supply, etc. Annex III. Point 2.

[30] See Article 13.

[31] Article 27.1.f) AI Act.

pact assessment under the AIA will be complementary to the data protection impact assessment. Interestingly, the European Parliament called for such a data protection impact assessment to be published, but this proposal was not successful[32].

After the impact assessment has been carried out, the deployer shall notify the competent market surveillance authority of the results of the impact assessment. This notification shall not be made in exceptional cases where the market surveillance authority has authorised the use of the AI system even though the AI system has not undergone a conformity assessment process[33].

In our view, the design of this impact assessment has two clear objectives.

On the one hand, the potential risks that this AI system may generate for the fundamental rights of individuals are specifically materialised, as well as the measures to mitigate them in the specific context where this AI system will be used. Recall that the provider will already have implemented a risk management system and will have taken into account the potential impact on fundamental rights[34]. This risk system developed by the provider will in many cases be the essential basis for the development of the impact assessment. This will especially be the case when the AI system has been developed specifically for the deployer and the provider knows in advance the potential uses and even the specific target groups for which the system will be used. However, the impact assessment will especially bring added value when such an AI system is implemented by different deployers who, while taking into account the risk system presented by the provider, will have to adapt it to their specific context while respecting the intended purpose of the AI system.

On the other hand, the impact assessment will also be important for the information that the deployer will provide to the competent market surveillance authorities about the system. This is relevant because it should be recalled that most of the AI systems that are regulated in Annex III provide for self-assessment of conformity, which means that beyond the internal control of the provider, there is no body prior to the deployment of the AI system that assesses in any way the conformity of the AI system. Through the system

[32] Amendment 419 of the European Parliament resolution of 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislation.

[33] In exceptional cases, the AI Act authorises market surveillance authorities to use high-risk AI systems despite the fact that they have not passed a conformity assessment process. Article 46.1 AI Act.

[34] Article 9.2.a) AI Act.

registration obligations in the Act and the impact assessment notification obligation, market surveillance authorities at least have on their radar the high-risk AI systems in use, as well as essential information about these systems and the entities using them.

3. The authorised representative and his obligations

An authorised representative is defined by the AIA as any natural or legal person located or established in the EU who has received and accepted a written mandate from an AI system provider to fulfil the obligations and carry out the procedures set out in this European Standard on behalf of that provider[35].

Before a provider established outside the EU wants to market its AI system in the EU, it must appoint an authorised representative established in the EU by means of a written mandate. In this regard, it should be noted that not all European product legislation makes this mandatory[36], and some do not even mention the authorised representative at all.

The mandate established between the provider and the authorised representative is essential when assessing the tasks to be performed by the authorised representative.

Firstly, these tasks enable the authorised representative to verify that the provider has carried out the relevant conformity assessment as well as the technical documentation and the declaration of conformity of the AI system.

Secondly, they shall keep for at least 10 years the contact details of the provider a copy of the declaration of conformity, the technical documentation of the AI system and, where applicable, the certificate issued by the notified body when it has been involved in the conformity assessment process[37].

Thirdly, they must either register the AI system and information about it in the EU database established by the AIA[38], or, if the AI system is already

[35] Article 3.5 of the AI Act.

[36] This is the case for example in legislation on products such as recreational craft or toys. Directive 2013/53/EU of the European Parliament and of the Council of 20 November 2013 relating to recreational craft and personal watercraft and repealing Directive 94/25/EC. Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys.

[37] Depending on the type of AI system, conformity assessment may or may not require the involvement of a Notified Body. See Article 43 of the AI Act and the chapter in this collective work that discusses conformity assessment.

[38] Registration obligations are foreseen for all AI systems in Annex III, except for the AI systems in point 2 of Annex III (critical infrastructure). Article 49.1. AI Act.

registered by the provider, ensure that the information that is incorporated in this register complies with the requirements[39].

Fourthly, authorised representatives are obliged, upon request of the market surveillance authorities, to provide the information they hold on the AI system for the purpose of demonstrating the compliance of the AI system. In particular, it is expressly stated that access to the log files being automatically generated by the AI system must be provided where such files are under the control of the provider. This makes sense, as the provider is located outside the EU. In addition, providers must cooperate in any actions taken by the latter to reduce or mitigate the risks that the AI system may present.

If the authorised representative has reason to believe that the provider is failing to fulfil the obligations laid down in the AIA, he shall terminate the mandate and inform the competent market surveillance authority and, where applicable, the notified body that assessed the conformity of the AI system accordingly.

4. The importer and his obligations

The importer is any natural or legal person located or established in the EU who places on the market an AI system bearing the name or trademark of a natural or legal person established in a third country[40].

As will now be seen, the functions assigned to importers in the Act reflect the fact that this operator is not to be considered as a mere reseller of AI systems, but plays a crucial role in ensuring the conformity of imported products[41].

Firstly, the importer, before placing an AI system on the market, must verify that the provider has carried out a series of compulsory actions on the system, such as: having carried out the relevant conformity assessment, having drawn up the technical documentation, verifying that the system bears the CE marking and the declaration of conformity, and that an authorised representative has been appointed. As can be seen, this is a whole series of obligations that the AIA requires of any provider of an AI system before they wish to place it on the market. In these cases, given that the introduction is carried out by the importer, the latter must corroborate that the system is fit for purpose.

[39] The information to be entered into the EU database (Article 71) is listed in Section A of Annex VIII of the AI Act.

[40] Article 3.6. AIA.

[41] Blue Guide on the implementation of the 2022 EU product legislation. p. 33.

If the importer finds that the AI system is not in conformity with the law, he shall not place it on the market until conformity of the AI system has been achieved. It will be necessary for the importer to contact the provider to clarify any doubts about the conformity of the product.

Secondly, if importers have reason to believe that a system presents a risk[42], they should communicate this to the system provider, authorised representatives, and relevant market surveillance authorities.

Thirdly, importers shall ensure that the storage or transport conditions of AI systems do not affect their essential requirements[43]. In addition, where appropriate, the manufacturer's name or trade name and contact address must be indicated on the documentation or packaging.

Fourthly, importers must keep for at least 10 years after the AI system has been placed on the market its instructions for use, the declaration of conformity and, where applicable, the copy of the certificate of conformity issued by a notified body[44].

Fifth, importers are obliged to provide information they have on the AI system to market surveillance authorities upon request and to cooperate in all actions taken by the latter to reduce or mitigate the risks that the AI system may present.

5. The distributor and his obligations

The distributor is any natural or legal person in the supply chain, other than the provider or the importer, who makes an AI system available on the EU market[45].

The AIA establishes a number of obligations on distributors.

Firstly, before making an AI system available on the market, distributors shall verify that the system bears the required CE marking, a copy of the declaration of conformity[46], the instructions for use, and the importer's or

[42] The concept of risk is defined in Article 79.1 of the AI Act.

[43] These requirements are set out in Articles 8 to 15 of the AI Act. They are: Risk management system, data and data governance, technical documentation, record keeping, transparency and communication of information, human oversight, accuracy, robustness and cybersecurity.

[44] Depending on the type of AI system, conformity assessment may or may not require the involvement of a Notified Body. See Article 43 of the AI Act and the chapter in this collective work that discusses conformity assessment.

[45] Article 3.7 of the AI Act.

[46] On several occasions, distributors of products have been sanctioned for placing on the market products that did not have the CE marking or the declaration of conformity. NA of 20 May 2010. (JUR 2010\182746).

provider's trade name. In the latter case, the provider must also have supplied the distributor with the quality management system.

If, on the basis of the information provided, the distributor considers that the AI system is not in conformity with the essential requirements of the AIA, the distributor shall not make the system available on the market until the system has been brought into conformity. If the distributor also identifies that the system presents a risk[47], the distributor shall inform the provider or the importer. In other words, the distributor should not supply an AI system if he knows or assumes, on the basis of the information in his possession and his professional experience, that it is not in conformity with the AIA[48].

Secondly, once the AI system is placed on the market, if the distributor considers that the AI system is not in conformity with the essential requirements of the AIA, it shall take appropriate action to correct the situation. These actions may include recalling the system, taking back the system, or ensuring that the provider or importer takes appropriate measures to reverse the problem.

In addition, if the distributor who made the system available on the market identifies a risk, he shall notify the provider or the importer and the market surveillance authorities of the market where he made the AI system available, informing them of the risk and of the corrective measures taken.

Thirdly, distributors, as well as importers, must ensure that the storage or transport conditions of the AI system do not compromise the essential requirements of the AI system. Therefore, the person(s) in charge of distribution conditions should take the necessary measures to protect the conformity of the AI system[49].

Fourthly, distributors are obliged to provide the information they have on the AI system to the market surveillance authorities upon request and to cooperate in all actions carried out by the latter to reduce or mitigate the risks that the AI system may present. Note that the role of the distributor differs, among other things, from that of the provider or importer in that the former has much less information on the AI system than the latter, but his role is also relevant throughout the entire AI system supply chain.

[47] The concept of risk is defined in Article 79.1 of the AI Act.

[48] Blue Guide on the implementation of the European 2022 product legislation. P.35.

[49] Blue Guide on the implementation of the 2022 EU product legislation. P.36.

Obligation	Authorised representative	Importer	Distributor
Ensuring that the AI system complies with the Act	x	x	x
Do not place on the market if there are doubts		x	x
Report in case of risks in the system		x	x
Indicate the name or trademark		x	
Preservation of documentation	x	x	
Duty to cooperate with the authorities	x	x	x
Registration obligations	x	x	

6. Possible alterations to operators' responsibilities

The AIA foresees different situations in which operators[50] present during the AI value chain other than the provider are considered as providers and assume the obligations required of the latter on the AI system[51]. These situations are:

Firstly, where such operators place their name or trademark on a high-risk AI system that has previously been placed on the market or put into service. In such cases, the provider and the operator may enter into contractual arrangements providing for the sharing of obligations in another way.

Secondly, where the operator substantially modifies a high-risk AI system that has been placed on the market or put into service, provided that such a system is still considered to be high risk after such substantial modification[52].

Thirdly, where the operator changes the intended purpose of the AI system in such a way that it is considered high risk when initially and without the change of purpose, such a system on the market was not considered high risk.

In all these cases, the original provider will no longer be considered as a provider for the purposes of the AIA. However, the initial provider will have to cooperate closely with the new provider and facilitate the necessary information to enable the new provider to comply with the obligations required by

[50] Any distributor, importer, deployer, or third party.

[51] See the obligations required of the provider in Article 16 of the AI Act.

[52] Substantial modification means "*a change to an AI system after its placing on the market or putting into service which is not foreseen or planned in the initial conformity assessment carried out by the provider and as a result of which the compliance of the AI system with the requirements set out in Chapter III, Section 2 is affected or results in a modification to the intended purpose for which the AI system has been assessed*". Article 3.23 AI Act.

the Act[53]. It is therefore a forward-looking obligation which does not depend *a priori* on the addressee of the obligation (initial provider) but on subsequent operators who may be considered as providers.

Such cooperation obligations shall not be required from the original provider where the latter has made it clear that its AI system which was not initially high risk should not be altered in such a way as to be considered as such.

In turn, in cases where a high-risk AI system that is a safety component of products covered by EU harmonisation legislation, the manufacturer of the product will be considered a provider when either the system is placed on the market under the name or trademark of the product manufacturer, or the system is put into service under the name or trademark of the product manufacturer after the product has been placed on the market.

Finally, where a third party provides a provider of a high-risk AI system with tools, services, components or processes that integrate the high-risk AI system, that third party and the provider must enter into a written agreement specifying what information and documentation, if any, is required to enable the provider to comply with the AIA. This obligation shall not apply to third parties that make tools, services or components other than general-purpose AI models available to the public under a free and open licence.

IV. The notifying authorities

1. Concept of notifying authority

The AIA obliges Member States to designate different authorities in order to ensure compliance. These authorities include the notifying authority.

According to Article 3.19 of the AIA, the notifying authority is responsible for setting up and carrying out the necessary procedures for the assessment, designation and notification of conformity assessment bodies, as well as for their monitoring[54].

The concept of notifying authority is not new to the AIA; other European texts already mention this concept. The AIA follows the structure of the New Legislative Framework (NLF). The NLF is made up of several European texts establishing common bases for the marketing, evaluation and

[53] Article 25.2. AI Act.

[54] Article 3.19 AI Act.

surveillance of products in the European Union[55]. All these texts provide for a notifying authority.

Each Member State will decide whether to appoint one or more notifying authorities, with at least one per country[56]. To date, as a general rule, the national authorities that have been designated to carry out these activities under other legislation regulating the marketing and supervision of products that also follow the NLF are Directorates General or Sub-Directorates General integrated within a given Ministry[57]. This is the case for example for products such as toys, lifts, radio equipment, among others[58].

Product	Notifying Authority
Machines	SG Quality and industrial safety (Ministry of Industry)
Gaseous fuel-burning appliances	
Pressure equipment	
Cableway installations	
Individual protection	
Lifts	
Protection for use in potentially explosive atmospheres	
Radio equipment	State Secretariat for Telecommunications and Digital Infrastructures (Ministry of Digital Transformation)
Toys	Directorate-General for Consumption (Ministry of Consumer Affairs)
Recreational craft	Directorate-General for Merchant Navy (Ministry of Transport)
Medical devices	Ministry of Health

[55] The three legal texts that make up the New Legislative Framework are: Regulation (EC) No 765/2008 of the European Parliament and of the Council setting out the requirements for accreditation and market surveillance of products; Decision No 768/2008/EC of the European Parliament and of the Council on a common framework for the marketing of products and; Regulation (EU) 2019/1020 of the European Parliament and of the Council on market surveillance and product conformity.

[56] Articles 28.1 and 70.1 of the AI Act.

[57] The full list of notifying authorities can be found at

https://webgate.ec.europa.eu/single-market-compliance-space/#/notified-bodies/notifying-authorities?filter=countryId:724

[58] These products are, in particular, machines, toys, lifts, equipment and protective systems intended for use in potentially explosive atmospheres, radio equipment, pressure equipment, recreational craft equipment, cableway installations, equipment burning gaseous fuels, medical devices and *in vitro* diagnostic medical devices. Recital 50 of the AIA.

Taking into account the above, the notifying authority in Spain for compliance with the AIA is likely to be the Secretary of State for the Digitalisation of Artificial Intelligence or an administrative body under it[59]. However, there may be other notifying authorities if deemed necessary by Spain in specific areas such as justice or law enforcement, among others.

2. Activities of the notifying authority

The activities assigned to these authorities focus on ensuring that a public or private entity has sufficient human, organisational, and technical means to be able to perform conformity assessment of AI systems covered by the AIA. These entities are known as conformity assessment bodies[60].

Before a conformity assessment body can perform conformity assessments of AI systems, the notifying authority shall first verify that the conformity assessment body is capable of carrying out conformity assessments of AI systems. Once the notifying authority ascertains that that conformity assessment body fulfils the requirements to be able to conduct conformity assessments of AI systems[61], the notifying authority shall "notify" the European Commission and the other Member States of that situation. From that moment on, that body will be a "notified body" entitled to carry out conformity assessments of AI systems under the AIA.

Among the activities assigned to the notifying authority, assessment and supervision may be entrusted to a national accreditation body[62], in the case of Spain, this accreditation body is currently ENAC[63].

[59] Within the Secretariat of State for Digitalisation and Artificial Intelligence is the Directorate General for Digitalisation and Artificial Intelligence and within the latter the Subdirectorate General for Artificial Intelligence and Digital Enabling Technologies. Royal Decree 210/2024, of 27 February, establishing the basic organisational structure of the Ministry for Digital Transformation and the Civil Service.

[60] Conformity Assessment Body means "*an independent body that performs third-party conformity assessment activities, including testing, certification and inspection*". Article 3.21. AI Act.

[61] The notification process for conformity assessment bodies is found in Articles 29 and 30 of the Artificial Intelligence Act. There is a section of this work that specifically discusses it.

[62] National Accreditation Body: "the only body in a Member State with public authority to carry out accreditation". Article 2.11. Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93.

[63] Royal Decree 1715/2010 of 17 December 2010 designating the National Accreditation Body (ENAC) as the national accreditation body in accordance with Regulation (EC) No 765/2008 of the European Parliament and of the Council of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products and repealing Regulation (EEC) No 339/93.

In carrying out the assigned activities, the notifying authority may not perform any activities that notified bodies perform, or consultancy services on a commercial or competitive basis. Furthermore, they shall avoid any conflict of interest that may arise between the conformity assessment bodies and the notifying authority. This is to ensure that the assessment and notification of conformity assessment bodies is carried out as impartially and objectively as possible by the notifying authority.

The requirement of independence and objectivity required of these authorities in their activities was not contemplated in the first version of the AIA[64], but has been incorporated in the various changes that have been introduced. These requirements of impartiality and objectivity are fully applicable with regard to the potential conformity assessment bodies that they will assess, whether they are public or private. In this sense, most of the notified bodies that have been notified to carry out conformity assessments of other products are private, however, nothing prevents it from being a public body, as is the case with the verification processes for medical devices[65].

Furthermore, the notifying authority must be organised in such a way that decisions relating to notification are taken by competent persons other than those who carried out the assessment of conformity assessment bodies. These persons should have sufficient competence to perform the tasks assigned to them adequately, including, where appropriate, expertise in areas such as information technology, Artificial Intelligence and fundamental rights. The latter is essential, as unlike other product standards that focus on reducing or mitigating risks related to human health and safety, in the case of AI systems, the potential fundamental rights that may be affected must also be taken into account.

Activities of notifying authorities[66]	**Articulated**
General regulation of Notifying Authorities	Article 28
Conformity Assessment and Notification of Conformity Assessment Bodies	Article 29 and 30
Monitoring of notified bodies	Articles 33(4), 34(3), 36, 37, 38 and 45

[64] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislation. 21 April 2021.

[65] The National Centre for the Certification of Medical Devices is the only notified body in Spain to carry out the conformity assessment of medical devices in accordance with Regulation 2017/745. Article 35.bis. Royal Decree 1275/2011, of 16 September, creating the State Agency "Agencia Española de Medicamentos y Productos Sanitarios" and approving its Statute.

[66] These activities are discussed in more detail in this book in the section on notified bodies.

V. Measures targeting small-scale providers and users

There is no doubt that the AIA aims to reduce or mitigate the risks that can and will be generated by the use of Artificial Intelligence systems. To achieve this objective, the different operators present during the lifecycle of AI systems must comply with all the obligations imposed on them by this regulation.

The implementation and adaptation to this standard will be easier for those private entities that have more staff and resources. These entities are also the ones that in many cases are developing techniques to comply with the regulatory requirements. For example, large companies in the sector, such as Microsoft, IBM, Google and others, are constantly investing heavily in implementing or facilitating so-called explainable Artificial Intelligence. In other words, their regulatory compliance techniques will set the guidelines to be followed by smaller entities.

To ensure that the innumerable regulatory requirements of this standard do not stifle companies that do not have the resources to adapt to it, the AIA provides a range of measures aimed at providers and deployers of AI systems that are SMEs and start-ups.

These measures are essentially binding on three parties: the Member States, the notified bodies, and the AI Office.

As far as *Member States* are concerned, first of all, they should prioritise the participation of SMEs in the controlled test sites they intend to carry out. This is contemplated, for example, in Spain's controlled test environment[67]. In this sense, the AIA itself establishes that one of the purposes of these controlled test sites will be to facilitate and accelerate access to the EU market for AI systems developed by SMEs[68].

Secondly, appropriate advisory channels should be established to help these companies to properly implement the AIA, as well as training on this same standard and its implementation. Some of these functions are already attributed to AESIA, regardless of the size or type of company[69].

[67] Article 8.1.j). Royal Decree 817/2023 of 8 November establishing a controlled test environment for testing compliance with the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence. This is also provided for in the AESIA Statute. Article 25.a). 4º. Royal Decree 729/2023, of 22 August, approving the Statute of the Spanish Artificial Intelligence Supervisory Agency.

[68] Article 57.9.e) AI Regulation.

[69] Article 4.3. a) and b). Royal Decree 729/2023 of 22 August, approving the Statute of the Spanish Agency for Supervision of Artificial Intelligence.

Thirdly, Member States should encourage the participation of SMEs in the standardisation development process. This is essential. Recall that one way to make it easier for product manufacturers to comply with a European directive or regulation applicable to that product is to use harmonised standards developed by European standardisation bodies[70]. Thus, although harmonised standards are not mandatory, their use gives presumption of conformity to products that have been designed using them as a reference, which is why manufacturers normally refer to them[71]. It will be essential for the SME and start-up sector to play an important role in the process of drawing up these harmonised standards[72].

As for the *notified bodies*, firstly, they are obliged to establish different fees depending on the type and size of the company for the services provided during the conformity assessment of AI systems regulated in the AIA. It has been estimated that a AIA conformity assessment process can cost between €16,800 and €23,000 for entities intending to place their AI systems on the market[73].

Secondly, the conformity assessment process related to the submission of technical documentation to be provided by SMEs to notified bodies is also streamlined. Thus, the European Commission will develop a simplified form to assist SMEs in documenting the technical documentation to be submitted to Notified Bodies when verifying that their AI system complies with the AIA[74].

Finally, the *AI Office* will also develop actions to promote the correct adaptation of SMEs to the AIA. These actions include raising awareness of regulatory compliance, the creation of a single information platform on this legal text, as well as the design of standardised models to help implement the different obligations set out in the AIA.

[70] These European standardisation bodies are CEN, CENELEC, and ETSI.

[71] Álvarez García, V and Tahirí Moreno, J., "La regulación de la inteligencia artificial en Europa a través de la técnica armonizadora del nuevo enfoque". *Revista General de Derecho Administrativo*, núm 63 (2023).

[72] McFadden/Jones/Taylor/Osborn, *Harmonising Artificial Intelligence: The role of standards in the EU AI Regulation*, Oxford Commission on AI & Good Governance (2021). P.20.

[73] European Commission. *Study to support an impact assessment of regulatory requirements for Artificial Intelligence in Europe.* 2021. P.12.

[74] Article 11.1 European Regulation on Artificial Intelligence.

Obliged party	Obligation	Article
Member State	Prioritisation of test pilot participation Individualised advice Specific training Participation in the standardisation process	62.1.
	Prioritisation of test pilot participation	57.9.e)
Notified Body	Adapted fees on conformity assessments	62.2
	Simplified forms	11.1
AI Office	Awareness-raising on the implementation of the Act.. Creation of an information platform Design of standardised models	62.3

VI. Notification of serious incidents

1.Concept of serious incident

According to Article 3.49 of the AIA, a serious incident is an incident or malfunction of an AI system which directly or indirectly results in: the death of a person or serious damage to the health of a person, serious disruption of critical infrastructure operations, breach of an obligation under Union law intended to protect fundamental rights, or serious damage to property or the environment.

As can be seen, the consideration of a serious incident is designed for the most relevant impacts that an AI system can generate in the sphere of people, property or the environment. In this sense, AI systems are implemented in different healthcare products used for disease detection or patient operations, in critical infrastructures such as water pipelines or power line deployment, or in the handling of huge amounts of personal data. In all these cases, an incident in the operation of an AI system will be considered serious.

In all such cases, the notification shall aim to reduce the risk that such a serious incident may recur or, in the event of a recurrence, to mitigate any damage that may have been caused.

We consider it relevant to analyse the cases in which an incident is considered serious when it involves a breach of obligations under Union law aimed at protecting fundamental rights[75].

There are two cumulative conditions that are foreseen in order to consider this serious incident, on the one hand, that it generates a breach of an obligation under EU law, and on the other hand, that this breach derived from the norm has the objective of protecting a fundamental right.

As regards the obligation deriving from EU law, we must include any obligation that is recognised in the different regulatory texts provided for in the EU legal system, as well as the national texts that may have been enacted by virtue of this European legislation. Take, for example, a Directive to be transposed or a European regulation that requires the cooperation of the Member States in order to develop certain elements of it. Fundamental rights, in turn, are to be understood as all the rights recognised in the Charter of Fundamental Rights of the European Union[76].

Ultimately, any incident in an AI system that results in a breach of an obligation recognised in EU law that affects a fundamental right recognised in the Charter will be considered a serious incident.

2. Who, before whom, and when must it be notified?

The provider of the AI system that experienced the serious incident has a responsibility to notify the market surveillance authorities of the Member States where the incident occurred[77]. However, where the incident has been detected by the deployer, the deployer shall notify the provider, and then the importer or distributor and the relevant market surveillance authority[78].

Note that the number of market surveillance authorities to be notified of the incident will vary depending on the number of Member States in which the incident may have occurred and the type of AI system that may have generated the incident. In this respect, the AIA provides for different market surveillance authorities for the different types of AI systems that it regulates[79].

As a general rule, the provider or, where applicable, the person responsible for the deployment, has a maximum of 15 days to report the incident

[75] Article 3.49(c) of the AI Act.

[76] We interpret this in the light of recitals 1, 2 and 48 of the Artificial Intelligence Regulation.

[77] Article 73.1. AI Act.

[78] Article 28.5. AI Act.

[79] For example, this is the case for AI systems intended for use in the banking sector, the judiciary or AI systems used for law enforcement purposes, among others. Article 74 AI Act.

from the time it occurred. However, this 15-day period will be reduced in certain circumstances.

In the first instance, notification shall be made as soon as the provider has established a causal link or a reasonable possibility of such a link between the AI system and the serious incident.

Secondly, the notification shall be made the day after the incident when the incident results in a serious and irreversible disruption to the management or operation of critical infrastructure or such an incident results in a widespread breach[80] , i.e., an act or omission contrary to EU law that affects or is likely to affect a group of persons in several Member States[81].

Third, when the incident results in the death of a person, the notification shall be made as soon as a causal link between the incident and the operation of the AI system is established. In any case, the notification shall not be postponed beyond 10 days after the occurrence of the incident.

Type of incident	Deadline	
	Minimum	Maximum
General rule	As soon as the causal link is known	15
Critical infrastructure or widespread infringement	Next day	
Death of a person	As soon as the causal link is known	10

In addition to the maximum reporting deadlines, the AIA provides for two cases in which, depending on the type of AI system, incident reporting is reduced only to certain cases.

On the one hand, for high-risk AI systems that are products or safety components of medical devices or in vitro diagnostic products[82], the notification of serious incidents will be limited to cases in which such incidents have led to a breach of obligations under EU law aimed at protecting fundamental rights[83]. We understand that this provision is due to the fact that the laws of these products already provide for their own notifications of serious inci-

[80] The concept of critical infrastructure is defined in Article 3.62 of the AI Act.

[81] The concept of a generalised infringement is described in Article 3.61 of the AI Act.

[82] These products are regulated in Regulation 2017/745 (medical devices) and Regulation 2017/746 (in vitro diagnostic medical devices).

[83] Article 73.11 of the AI Act.

dents, which have a very similar structure to that laid down in the AIA, but adapted to the reality of medical devices[84].

On the other hand, for high-risk AI systems listed in Annex III of the AIA that are subject to Union legislative instruments providing for equivalent obligations on serious incident reporting, the reporting of such incidents shall be limited to the scenario indicated above, i.e., where such incidents have led to a breach of obligations under Union law aimed at protecting fundamental rights[85].

3. Proceedings following the notification of the incident

Once the serious incident has been notified to the competent Market Surveillance Authority, the AIA provides for different actions to be taken by providers and Market Surveillance Authorities.

On the *providers*' side, they will conduct the necessary investigations to clarify the incident that occurred. We consider these investigations to have been initiated prior to the notification of the incident when the possible causal link between the incident and the failure of the AI system has been investigated at an early stage.

In turn, providers should carry out a risk assessment of such an incident and the corrective measures to reduce or mitigate them[86]. It is possible that the different potential serious incidents that could occur have been contemplated in the risk system that every provider is required to develop for its AI system[87]. If such incidents were foreseen in that risk system, the provider or deployer can implement the measures foreseen in it.

In addition, they shall cooperate with the different authorities and, where appropriate, with the notified body that has assessed the conformity of their

[84] See Article 2.65 on the definition of a serious incident and Article 87(1) on the reporting of such incidents in Regulation 2017/745 (medical devices). In turn, see Article 2.68 on the definition of serious incident and Article 82 on serious incident reporting in Regulation 2017/746 (in vitro diagnostic medical devices).

[85] In the initial versions of the AI Act this provision was specifically designed for those cases where a financial institution used an AI system for the purpose of assessing the financial solvency of persons (Annex III. Point 5 b) of the AI Regulation). However, in the latest known version this express mention has been removed. Nevertheless, we continue to believe that this perception is essentially designed for these cases. See the initial version of Article 62(3) of the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislation. Resolution of 21 April 2021.

[86] Article 73.7 AI Act.

[87] Article 9. AI Act.

AI system[88]. They shall not make any modifications to the system that could have an impact on future investigations into the causes of such an incident.

Market surveillance authorities[89], once they have been notified of such a serious incident, will, within a maximum of 7 days, take different measures depending on the seriousness of the incident. These measures may range from a ban on the sale of the AI systems to the withdrawal or recall of the systems[90]. In addition, they must inform the European Commission if the measures they intend to take go beyond the borders of the Member State where the market surveillance authority exercises its competence. This is often the case with AI systems, which are often designed to offer their services in various Member States.

Irrespective of whether or not the above measures are taken, the market surveillance authority must report the serious incident to the European Commission through the rapid information exchange system provided for in the European Market Surveillance Regulation[91].

On the other hand, when a market surveillance authority is notified of a serious incident involving a breach of an obligation under European law to protect a fundamental right, it must inform the competent authority responsible for monitoring and enforcing compliance with the fundamental right affected by the incident[92]. Such authorities include, for example, those in charge of monitoring and supervising personal data protection rules.

VII. Conclusions

1. The AIA defines and sets out the different roles and obligations of the different operators involved in the value chain of AI systems.

2. These obligations are mainly addressed to providers and, to a lesser extent, to deployers.

[88] Depending on the type of AI system, conformity assessment may or may not require the involvement of a Notified Body. See Article 43 of the AI Regulation and the chapter of this collective work that discusses conformity assessment.

[89] Articles 73.8, 9 and 12 of the AI Act.

[90] Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and product conformity and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011.

[91] Article 73(12) of the AI Act and Article 20 of Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and product conformity and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011.

[92] Article 73.8 77.1 of the AI Act.

3. Importers, distributors, and authorised representatives are also assigned certain responsibilities specific to the functions they perform.

4. Any operator other than the provider may become a provider when it takes certain actions in respect of the AI system.

5. Notifying authorities have the primary role to monitor, assess, and designate conformity assessment bodies. Once a conformity assessment body has been notified to carry out such assessments under the AIA by a notifying authority, it shall be considered a notified body. Notifying authorities shall have the necessary human and technical resources to carry out these activities.

6. The AIA offers a range of measures specifically designed for SMEs who either provide or oversee the deployment of AI systems. The objective is clear: due to the complexity that the adaptation and integration of the requirements demanded by this standard may entail for them, certain subjects are obliged to implement these support measures to help them adapt to the standard correctly.

7. The AIA provides for a notification process for serious incidents that may occur in an AI system. This is intended to establish a procedure to mitigate or reduce as far as possible the effects that such incidents have caused or may cause in the future on other AI systems with the same characteristics.

SUBJECTS AND ACTORS IN CONFORMITY ASSESSMENTS (NOTIFIED BODIES)

Ignacio Alamillo Domingo
PhD in Law[1]

I. Introduction

High-risk AI systems are subject to mandatory conformity assessment in accordance with Article 16(f) of the AIA, which refers to Article 43 of the Regulation itself, and in certain cases this assessment must be carried out by a notified body for this purpose.

This chapter presents the legal regime applicable to such Notified Bodies for the performance of conformity assessment activities,[2] addressing mainly the content of Articles 29-39 and 44-46 of the AIA.

The AIA defines a notified body as a conformity assessment body notified under the Regulation and other relevant Union harmonisation legislation (Article 3.22), where a conformity assessment body is a body that performs third party conformity assessment activities, including testing, certification and inspection (Article 3.21 of the AIA).

While we are dealing with a specific regime established by the AIA, Recital 46 of the AIA clarifies that, as part of the Union harmonisation legislation,

[1] EID, trust and security legal freak. With a PhD in Law about eIDAS. CISA, CISM, CDPSE.

[2] The literature on this subject is strikingly scarce, I would like to point out the most relevant studies: De Lucia L., "One and Triune – Mutual Recognition and the Circulation of Goods in the EU", *Review of European Administrative Law*, 13 (3), 2020, pp. 7-35; De Vries S., Kanevskaia O., De Jager R., "Internal Market 3.0: The Old "New Approach" for Harmonising AI Regulation", *European Papers – A Journal on Law and Integration*, 8 (2), 2023, pp. 583-610; Demetzou, K., "Introduction to the conformity assessment under the draft EU AI Act, and how it compares to DPIAs", Future of Privacy Forum, August 12, 2022; Galland J.-P., "The difficulties of regulating markets and risks in Europe through notified bodies", *European Journal of Risk Regulation*, 4 (3), 2013, pp. 365-373 and "La difficile construction d'une expertise européenne indépendante", *Revue d'anthropologie des connaissances*, 7-1, 2013; Holder C., Hawes C., Hatzel, J., "The Commission's proposed Artificial Intelligence Regulation", *Computer and Telecommunications Law Review*, 27 (5), 2021, pp. 130-134 and Lohbeck, D., "Chapter 4 – Notified Bodies and Certification", *CE Marking Handbook*, Newnes, 1998, pp. 53-63; Tricker, R., "2 – Structure of new approach directives", *CE Conformity Marking*, Butterworth-Heinemann, 2000, pp. 46-54; Veale, M., Borgesius, F.Z., "Demystifying the Draft EU Artificial Intelligence Act. Analysing the good, the bad, and the unclear elements of the proposed approach", *Computer Law Review International*, 4/2021.

the rules applicable to the placing on the market, putting into service and use of high-risk AI systems should be established in a manner consistent with the "New legislative framework for the marketing of products", contained in Regulation (EC) No 765/2008 of the European Parliament and of the Council setting out the requirements for accreditation and market surveillance of products, Decision No 768/2008/EC of the European Parliament and of the Council on a common framework for the marketing of products, Decision No 768/2008/ EC of the European Parliament and of the Council on a common framework for the market surveillance and conformity of products, and Regulation (EU) 2019/1020 of the European Parliament and of the Council on market surveillance and conformity of products.No 768/2008/EC of the European Parliament and of the Council on a common framework for the marketing of products and Regulation (EU) 2019/1020 of the European Parliament and of the Council on market surveillance and product conformity; the relevant rules of this framework will therefore be applicable in a supplementary manner, as described in the Commission Communication (2022/C 247/01) "Blue Guide" on the implementation of the European product legislation of 2022.

Conformity assessment refers to the process of demonstrating compliance with the requirements set out in Section 2 of Chapter III of the AIA, relating to high risk AI systems, to the detailed analysis of which we refer. This definition specifies the general definition contained in Article 2.12 of Regulation (EC) No 765/2008, namely the process of demonstrating whether specific requirements relating to a product, process, service, system, person or body are fulfilled.

In this regard, it is not superfluous to recall that, according to Article 6.1 of the AIA, an AI system is considered to be high risk when the AI system is intended to be used as a safety component of a product, or the AI system itself is a product, covered by Union harmonisation legislation listed in Annex I, provided that the product of which the AI system is the safety component, or the AI system itself as a product, is subject to a third party conformity assessment, with a view to the placing on the market and/or putting into service of that product in accordance with Union harmonisation legislation listed in Annex I.

Annex I includes both Union harmonisation legislation based on the New Legislative Framework, 12 legal acts, and other Union harmonisation legislation, 8 legal acts. In the 12 cases set out in Section A of Annex I, the provider shall carry out the relevant conformity assessment in accordance with the provisions of those legal acts.

These include AI systems in certain areas, such as biometrics, critical infrastructure, education, employment, personnel management and access to self-employment, access to and enjoyment of certain essential services and benefits, public and private, law enforcement, border control, migration and asylum, and the administration of justice and democratic processes.

Although conformity assessment is always mandatory for high-risk AI systems, only in some cases is the intervention of a notified body required:

- For high-risk AI systems covered by the legal acts listed in Section A of Annex I of the AIA, i.e. machinery, safety of toys, recreational craft and personal watercraft, lifts and safety components for lifts, equipment and protective systems intended for use in potentially explosive atmospheres, placing on the market of radio equipment, placing on the market of pressure equipment, cableway installations, personal protective equipment, gas appliances, medical devices and in vitro diagnostic medical devices.

- In the case of AI systems for remote biometric identification systems, AI systems intended to be used for biometric categorisation, based on sensitive or protected attributes or characteristics based on the inference of such attributes or characteristics, or AI systems intended for emotion recognition (Annex III.1). However, the intervention of the notified body is only necessary where no harmonised standards exist and no common specifications are available; or where the provider has not implemented or has only partially implemented the harmonised standard; or where the common specifications exist but the provider has not implemented them; or where one or more of the harmonised standards have been published with a restriction and only in the part of the standard that was restricted.

In addition, where the system is intended to be put into service by law enforcement, immigration or asylum authorities, as well as by EU institutions, bodies, offices or agencies, the market surveillance authority foreseen for this purpose in the AIA itself shall necessarily act as notified body.

II. The notification procedure

In order to act as a conformity assessment body for high risk AI systems, it is necessary to fulfil a number of requirements and to be notified as such to the European Commission and the other Member States by a national notifying authority, responsible for setting up and carrying out the necessary procedures for the assessment, designation and notification of conformity assessment bodies and their monitoring, without prejudice to the possibility for the assessment and monitoring of conformity assessment bodies to be carried out by a national accreditation body within the meaning of and in accordance with Regulation (EC) No 765/2008.

Without prejudice to the above, Article 39 of the IAR provides that conformity assessment bodies established under the law of a third country with which the Union has concluded an agreement may be authorised to carry out the activities of notified bodies under the IAR, provided that, according to the text agreed during the trilogues, such conformity assessment bodies comply with the requirements of Article 31 or ensure an equivalent level of compliance, which will be left to the agreement to be concluded between the Commission and each individual third country. This is already the case in other harmonisation legislation, including Australia, Canada, USA, Japan, New Zealand and Switzerland. It is conceivable that, at least in those cases where a high-risk AI system is integrated into a product subject to Notified Body conformity assessment, this provision may be particularly relevant.

In any case, the AIA does not exhaust all the aspects necessary for the implementation of the notification procedure, so it is foreseeable that the corresponding national legislation will specify and adapt the procedure, for example, for the purposes of the request for designation, or language requirements, as has occurred in other regulations, such as Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices. All actions relating to the notification procedure and its changes should therefore be carried out in accordance with the common administrative procedure rules, under the terms set out in the national rules, as is the case in other cases. Given that we will normally be dealing with applicants with the status of a legal person, this will be a fully electronic procedure. Competence should also be attributed to the corresponding body and issues such as the regime of administrative silence or appeals in the event of refusal of designation should be specified.

1. The application for designation

The notification procedure, regulated in Article 29 of the AIA, is initiated by the submission, by the candidate conformity assessment body, of an application for designation to the notifying authority of the Member State in which the candidate conformity assessment body is established.

The application for designation must be accompanied by a description of the conformity assessment activities to be performed, the conformity assessment module(s) and the types of CA systems for which the conformity assessment body claims to be competent, as well as an accreditation certificate, if available, issued by a national accreditation body, stating that the conformity assessment body fulfils the requirements set out in Article 31 of the AIA, to which reference will be made later.

In addition, the applicant body may add any valid document related to the existing designations of the applicant notified body under any other Union harmonisation legislation, which especially makes sense in cases where the high-risk AI system is incorporated as a component of a product subject to harmonisation legislation. This is not the first case where Union harmonisation legislation refers to software embedded in a product, certainly significantly in Directive 2006/42/EC on machinery or Directive 2014/53/EU on radio equipment.

In this regard, the AIA provides that bodies which have been notified under the legal acts provided for in Annex I of the AIA(harmonisation legislation under the New Legislative Framework or other harmonisation legislation) shall be empowered to monitor the conformity of high-risk AI systems with the requirements laid down for that purpose, but provided that the conformity of such notified bodies with the requirements laid down in Article 31(4), (9) and (10) has been assessed in the context of the notification procedure under those legal acts.

The conformity assessment activities provided for in Decision 768/2008 include calibration, testing, certification and inspection, while the conformity assessment modules are as follows: internal production control (plus supervised testing of products or plus supervised testing of products at random intervals), EC type examination, conformity to type based on internal production control (plus supervised testing of products or plus supervised testing of products at random intervals), EC type-examination, conformity to type based on internal production control (plus supervised testing of products or plus supervised testing of products at random intervals), conformity to type based on quality assurance of the production process, conformity to type based on product quality assurance, conformity to type based on product verification, conformity to type based on unit verification, and conformity to type based on full quality assurance.

The AIA does not directly target any of these modules, but they will be applicable depending on the provisions of the relevant harmonisation legislation, in cases where the conformity assessment of the high-risk AI system is carried out in conjunction with the conformity assessment of the product in which it is integrated. In these cases, in addition, the activities provided for in sections 4.3 to 4.5 and the fifth paragraph of section 4.6 of Annex VII of the AIA must also be carried out.

In the remaining cases, the conformity assessment procedure with intervention of the notified body, contained in Annex VII of the AIA, referring to the assessment of the quality management system and the assessment of the technical documentation, to the analysis of which we refer, shall simply apply.

Finally, where the applicant conformity assessment body has not previously been accredited, it shall provide the notifying authority with all documentary evidence necessary for the verification, recognition and regular monitoring of its compliance with the requirements laid down in Article 31, which will normally be the case where the notifying authority has chosen not to entrust the assessment and monitoring to a national accreditation body.

In the case of notified bodies designated under any other Union harmonisation legislation, all documents and certificates linked to such designations may be used to support their designation procedure. The Council's general approach has introduced an obligation for the notified body to update the documentation referred to in Article 29(2) and (3), if relevant changes occur, in order for the authority responsible for notified bodies to monitor and verify continued compliance with the requirements laid down in Article 31.

2. The notification procedure

The notification procedure as such is regulated in Article 30 of the IAM, certainly aligned with Article R23 of Decision (EU) No 768/2008, which initially mandates that notifying authorities may only notify conformity assessment bodies that have fulfilled the requirements set out in Article 31.

Recital 126, in the version proposed by the Commission, has been amended by both the European Parliament and the Council, whose wording has incorporated the need for notified bodies to comply with the relevant cybersecurity requirements in addition to the initially envisaged requirements of independence, competence and absence of conflict of interest. Moreover, the final wording of the Recital refers expressly to the use of the tool provided for in the aforementioned Article R23 of Decision (EU) No 768/2008.

It is therefore established that notifying authorities shall notify the Commission and the other Member States through the electronic notification tool developed and managed by the Commission, currently the NANDO (New Approach Notified and Designated Organisations) information system, accessible at https://webgate.ec.europa.eu/single-market-compliance-space/#/notified-bodies. At the proposal of the European Parliament, and somewhat in line with the general orientation of the Council, it has been clarified that each conformity body that has satisfied the requirements of Article 31 shall be notified.

With regard to the content of the notification, the original European Commission proposal was limited to detailed information on the conformity assessment activities, the conformity assessment module(s) and the Artificial Intelligence technologies concerned. However, both the European Parliament and the Council proposed to also include the relevant statement of competence of the conformity assessment body, which may be, where appropriate, the outcome of the accreditation procedure. And, at the proposal of the Council, the provision has also been added that, where a notification is not based on an accreditation certificate, the notifying authority shall provide documentary evidence attesting the competence of the conformity assessment body and the arrangements in place to ensure that the conformity assessment body will be monitored regularly and will continue to meet the requirements set out in Article 33. This is a requirement similar to that contained in other regulatory standards applicable to notified bodies, such as the aforementioned Regulation (EU) 2017/745.

Following notification through the NANDO information system, the conformity assessment body in question may perform the activities of a notified body only if no objections are raised by the Commission or the other Member States within two weeks of validation of a notification including an accreditation certificate, which will be two months where documentary evidence of the conformity assessment body's competence is provided. This is a different regime from the one initially proposed by the European Commission, which was of one month as a general rule and which was certainly not aligned with Article R23 of Decision (EU) No 768/2008.

On a proposal from the European Parliament, provision has been made, in the event of objections being raised, for the Commission to consult without delay the relevant Member States and the conformity assessment body; the Commission shall decide whether or not authorisation is justified, which decision shall be addressed to the Member State concerned and to the relevant conformity assessment body.

Finally, following a proposal by the Council, the provision contained in the Commission's original proposal that notifying authorities shall notify the Commission and the other Member States of any subsequent relevant changes to the notification has been transferred to Article 36 of the IAR.

3. Identification and publicity of notified bodies

Article 35 of the AIA, which has remained unchanged during its processing, deals with the identification of each notified body, which shall be assigned a unique number for all its conformity assessment activities, regardless of whether it has been notified under several Union acts. This number is managed in the aforementioned NANDO information system.

The same article also requires the Commission to make public the list of bodies notified under the Regulation, together with the identification numbers assigned to them and the activities for which they have been notified, and to ensure that the list is kept up to date, all of which is managed in the

NANDO information system, which is expected to be extended with the new harmonised legislation.

4. Changes to the notification

Article 36 of the AIA deals with the treatment of changes to notifications made, and is one of the articles concerning notified bodies that has undergone the most changes since the Commission's original proposal.

Firstly, and as mentioned above, a first heading has been added at the beginning of the article with the obligation for notifying authorities to notify the Commission and the other Member States of any subsequent changes to the notification that may be relevant, specifying that such notification of changes must be made through the electronic tool indicated in Article 30(2) of the IAR, i.e. the NANDO information system.

Secondly, and again at the proposal of the Council, Article 36(2) of the IAM clarifies that extensions to the scope of a notification already made shall entail a new notification procedure, including the corresponding application, which is logical since such an extension without prior control would allow for fraudulent circumvention of the law. For other changes to the notification, the procedures foreseen in the remaining paragraphs of the Article itself apply, namely cessation of activity by the notified body, or limitation, suspension or withdrawal of the notification, in case of non-compliance with the requirements applicable to the notified body.

The regime of cessation, contained in Article 36(3) of the IAR, introduced on a proposal from the Council, is applicable when a notified body decides to cease its conformity assessment activities, in which case it shall inform the notifying authority and the providers concerned as soon as possible and, in the case of a planned cessation, one year before the cessation of its activities. Where the notified body has ceased its activity, the notifying authority shall withdraw the designation.

As regards the regime of limitation, suspension or withdrawal of notification, the Commission proposal contained two headings, which have been significantly expanded during the course of the regulation, with particular emphasis on the Council, in particular in relation to the detail of the appropriate steps to be taken by the notifying authorities to ensure that the dossiers of that notified body are taken over by another notified body or are made available to the responsible notifying authorities on request. The text originally proposed by the Commission is very similar to that found in other harmonisation legislation, however in this case it was considered appropriate to increase the level of detail .

In this regard, according to Article 36(4) of the IAR, where a notifying authority has sufficient reason to consider that a notified body no longer meets the requirements laid down in Article 31 or that it is failing to fulfil its obligations, it shall promptly investigate the matter with the utmost dispatch, informing the notified body concerned of the objections raised and giving it the opportunity to make its views known, which administrative procedure shall be carried out in full compliance with the rules of common administrative procedure and with all the guarantees provided for by law.

In any case, where the notifying authority has come to the conclusion that the notified body no longer meets the requirements laid down in Article 31, or that it is failing to fulfil its obligations, it must necessarily take a decision restricting, suspending or withdrawing the notification, as appropriate, depending on the seriousness of the failure to meet those requirements or fulfil those obligations, and immediately inform the Commission and the other Member States thereof, obviously through the NANDO information system.

As already indicated, all these actions must be carried out in accordance with the rules of common administrative procedure, but must take into account the particularities provided for in the remaining sections of Article 36, which are obviously directly applicable .

Firstly, the new Article 36(5) of the AIA requires the notified body whose designation has been suspended, restricted or withdrawn in whole or in part, to inform the manufacturers concerned thereof within 10 days at the latest.

Secondly, according to the new Article 36(6) and (7) of the IAR, in case of restriction, suspension or withdrawal of a notification, the notifying authority shall take the following actions:

- take appropriate steps to ensure that the files of the notified body concerned are kept and are made available on request to the notifying authorities of other Member States and to market surveillance authorities.
- Assess the impact on the certificates issued by the notified body.
- Submit a report on its findings to the Commission and the other Member States within three months of notification of the changes introduced.
- Require the Notified Body to suspend or withdraw, within a reasonable period of time determined by the authority, all certificates that have been improperly issued to ensure the conformity of AI systems on the market.
- inform the Commission and the Member States of the licences whose suspension or withdrawal it has required.
- provide the competent national authorities of the Member State in which the provider has its registered office with all relevant information concerning the certificates for which it has requested suspension or withdrawal.

That competent authority shall take appropriate measures, where necessary, to avoid a potential risk to health, safety or fundamental rights.

Finally, the last paragraph of the new Article 36(9) of the IAM requires the national competent authority or the notified body assuming the functions of the notified body affected by the change of notification to immediately inform the Commission, the other Member States and the other notified bodies thereof.

5. Questioning the competence of notified bodies

Article 37 of the AIA deals with challenges to the competence of notified bodies, mandating in its first paragraph, as proposed by the European Parliament, that the Commission shall, where necessary, investigate all cases where there are grounds for doubting the competence of a notified body or the continued fulfilment by a notified body of the requirements laid down in Article 31 and its related responsibilities.

For this purpose, the notifying authority shall provide the Commission, on request, with all relevant information relating to the notification or the maintenance of the competence of the notified body concerned (section 2), which shall be treated confidentially, taking into account its sensitivity (section 3).

As a result of this investigation, where the Commission ascertains that a notified body does not meet or no longer meets the requirements for its notification, it shall inform the notifying Member State accordingly and request it to take the necessary corrective measures, including the suspension or withdrawal of notification if necessary, and, in the event of failure to act by the Member State, the Commission may itself take such corrective measures, as may be determined in an implementing act. This is provided for in heading 4, which has been significantly amended from the initial proposal, and which has adopted Parliament's position.

III. The performance of notified bodies

1. Requirements applicable to notified bodies

The important Article 31 of the AIA details the list of requirements for conformity assessment bodies eligible for notification, or their subsidiaries or subcontractors (Article 33), which include the following:

- be established under national law and have legal personality (heading 1, as proposed by the Council), without prejudice to the recognition of bodies established in third States.

- Meet the organisational, quality management, resource and process requirements necessary for the fulfilment of its tasks, as well as appropriate cybersecurity requirements (section 2). The addition of the cybersecurity requirements comes from the Council.

- Have an organisational structure, allocation of responsibilities, reporting lines and functioning that ensures confidence in the performance and results of the conformity assessment activities they carry out (section 3).

- be independent of the provider of a high-risk AI system in relation to which they perform conformity assessment activities, and of any other operator having an economic interest in the high-risk AI system under assessment, as well as of any competitor of the provider; this shall not preclude the use of assessed AI systems that are necessary for the operation of the conformity assessment body or the use of such systems for personal purposes (heading 4, final indent added on the basis of a Council proposal).

- Not be directly involved in the design, development, marketing or use of high-risk AI systems, nor represent parties engaged in such activities, or in any activity that may conflict with their independence of judgement or integrity in relation to the conformity assessment activities for which they are notified, which shall apply in particular to consultancy services (new heading 5, added on the basis of a Parliament proposal).

- Be organised and operated in such a way as to safeguard the independence, objectivity and impartiality of their activities by documenting and implementing a structure and procedures to safeguard impartiality and to promote and apply the principles of impartiality in all their organisational, personnel and evaluation activities (section 6).

- Have documented procedures to ensure that their personnel, committees, subsidiaries, subcontractors and any associated bodies or personnel of external bodies respect the confidentiality of the information which, in accordance with Article 78 of the AIA, comes into their possession in the course of carrying out conformity assessment activities, except where disclosure is required by law. Therefore, it is provided that the personnel of notified bodies shall be bound to observe professional secrecy with regard to all information obtained in carrying out their tasks, except in relation to the notifying authorities of the Member State in which their activities are carried out (heading 7, the only modification being, as proposed by the Council, the addition of the reference to Article 78 of the IAR).

- Have procedures for the conduct of activities that take due account of the size of an undertaking, the sector in which it operates, its structure and the degree of complexity of the AI system concerned (heading 8).

- take out appropriate liability insurance for their conformity assessment activities, unless liability is assumed by the Member State in which they are established in accordance with national law or that Member State is directly responsible for the conformity assessment (heading 9, slightly amended on a proposal from the Council to specify that the State concerned shall be the State of establishment of the notified body).

- be capable of carrying out all the tasks required of them under this Regulation with the highest degree of professional integrity and requisite competence in the specific field, whether these tasks are carried out by the notified bodies themselves or on their behalf and under their responsibility (section 10).

- Have sufficient internal competences to be able to effectively assess tasks performed by external parties on its behalf, for which the notified body shall have at its permanent disposal sufficient administrative, technical, legal and scientific staff with experience and expertise related to the relevant types of AI, data and data computing systems, as well as to the legally established requirements (heading 11, as proposed by the Council).

- Participate in the coordination activities referred to in Article 38 and participate in European standardisation organisations, either through direct involvement or representation, or by ensuring that they are aware of and keep up to date with the relevant standards (section 12).

- When subcontracting specific tasks connected with conformity assessment or having recourse to a subsidiary, ensure that the subcontractor or the subsidiary meets the requirements set out in Article 31 and inform the notifying authority accordingly (Article 33.1).

- Take full responsibility for the tasks performed by subcontractors or subsidiaries (Article 33.2).

- Subcontract or carry out conformity assessment activities through a subsidiary only with the agreement of the provider, and make publicly available a list of its subsidiaries (Article 33.3).

- Keep relevant documents concerning the assessment of the qualifications of the subcontractor or the subsidiary and the work carried out by them at the disposal of the notifying authority for a period of five years from the date of the termination of the subcontracting activity (Article 33(4), as proposed by the Council).

- be established under national law and have legal personality (heading 1, as proposed by the Council), without prejudice to the recognition of bodies established in third States.

- Meet the organisational, quality management, resource and process requirements necessary for the fulfilment of its tasks, as well as appropriate cybersecurity requirements (section 2). The addition of the cybersecurity requirements comes from the Council.

- Have an organisational structure, allocation of responsibilities, reporting lines and functioning that ensures confidence in the performance and results of the conformity assessment activities they carry out (section 3).

- be independent of the provider of a high-risk AI system in relation to which they perform conformity assessment activities, and of any other operator having an economic interest in the high-risk AI system under assessment, as well as of any competitor of the provider; this shall not preclude the use of assessed AI systems that are necessary for the operation of the conformity assessment body or the use of such systems for personal purposes (heading 4, final indent added on the basis of a Council proposal).

- Not be directly involved in the design, development, marketing or use of high-risk AI systems, nor represent parties engaged in such activities, or in any activity that may conflict with their independence of judgement or integrity in relation to the conformity assessment activities for which they are notified, which shall apply in particular to consultancy services (new heading 5, added on the basis of a Parliament proposal).

- be organised and operated in such a way as to safeguard the independence, objectivity and impartiality of their activities by documenting and implementing a structure and procedures to safeguard impartiality and to promote and apply the principles of impartiality in all their organisational, personnel and evaluation activities (section 6).

- Have documented procedures to ensure that their personnel, committees, subsidiaries, subcontractors and any associated bodies or personnel of external bodies respect the confidentiality of the information which, in accordance with Article 78 of the AIA, comes into their possession in the course of carrying out conformity assessment activities, except where disclosure is required by law. Therefore, it is provided that the personnel of notified bodies shall be bound to observe professional secrecy with regard to all information obtained in carrying out their tasks, except in relation to the notifying authorities of the Member State in which their activities are carried out (heading 7, the only modification being, as proposed by the Council, the addition of the reference to Article 78 of the AIA).

- Have procedures for the conduct of activities that take due account of the size of an undertaking, the sector in which it operates, its structure and the degree of complexity of the AI system concerned (heading 8).

- take out appropriate liability insurance for their conformity assessment activities, unless liability is assumed by the Member State in which they are established in accordance with national law or that Member State is directly responsible for the conformity assessment (point 9, slightly amended at the proposal of the Council to specify that the State concerned shall be the State of establishment of the notified body).

- be capable of carrying out all the tasks required of them under this Regulation with the highest degree of professional integrity and requisite competence in the specific field, whether these tasks are carried out by the notified bodies themselves or on their behalf and under their responsibility (section 10).

- Have sufficient internal competences to be able to effectively assess the tasks performed by external parties on its behalf, for which the notified body shall have at its permanent disposal sufficient administrative, technical, legal and scientific staff with experience and expertise related to the relevant types of AI, data and data computing systems, as well as to the legally established requirements (heading 11, as proposed by the Council).

- Participate in the coordination activities referred to in Article 38 and participate in European standardisation organisations, either by direct involvement or representation, or by ensuring that they are aware of and keep up to date with the relevant standards (section 12).

- When subcontracting specific tasks connected with conformity assessment or having recourse to a subsidiary, ensure that the subcontractor or the subsidiary meets the requirements set out in Article 31 and inform the notifying authority accordingly (Article 33.1).

- Take full responsibility for the tasks performed by subcontractors or subsidiaries (Article 33.2).

- Subcontract or carry out conformity assessment activities through a subsidiary only with the agreement of the provider, and make publicly available a list of its subsidiaries (Article 33.3).

- Keep relevant documents concerning the assessment of the qualifications of the subcontractor or the subsidiary and the work carried out by them at the disposal of the notifying authority for a period of five years from the date of the termination of the subcontracting activity (Article 33(4), as proposed by the Council).

- be established under national law and have legal personality (heading 1, as proposed by the Council), without prejudice to the recognition of bodies established in third States.

- Meet the organisational, quality management, resource and process requirements necessary for the fulfilment of its tasks, as well as appropriate cybersecurity requirements (section 2). The addition of the cybersecurity requirements comes from the Council.

- Have an organisational structure, allocation of responsibilities, reporting lines and functioning that ensures confidence in the performance and results of the conformity assessment activities they carry out (section 3).

- be independent of the provider of a high-risk AI system in relation to which they perform conformity assessment activities, and of any other operator having an economic interest in the high-risk AI system under assessment, as well as of any competitor of the provider; this shall not preclude the use of assessed AI systems that are necessary for the operation of the conformity assessment body or the use of such systems for personal purposes (heading 4, final indent added on the basis of a Council proposal).

- Not be directly involved in the design, development, marketing or use of high-risk AI systems, nor represent parties engaged in such activities, or in any activity that may conflict with their independence of judgement or integrity in relation to the conformity assessment activities for which they are notified, which shall apply in particular to consultancy services (new heading 5, added on the basis of a Parliament proposal).

- be organised and operated in such a way as to safeguard the independence, objectivity and impartiality of their activities by documenting and implementing a structure and procedures to safeguard impartiality and to promote and apply the principles of impartiality in all their organisational, personnel and evaluation activities (section 6).

- Have documented procedures to ensure that their personnel, committees, subsidiaries, subcontractors and any associated bodies or personnel of external bodies respect the confidentiality of the information which, in accordance with Article 78 of the AIA, comes into their possession in the course of carrying out conformity assessment activities, except where disclosure is required by law. Therefore, it is provided that the personnel of notified bodies shall be bound to observe professional secrecy with regard to all information obtained in carrying out their tasks, except in relation to the notifying authorities of the Member State in which their activities are carried out (heading 7, the only modification being, as proposed by the Council, the addition of the reference to Article 78 of the IAR).

- Have procedures for the conduct of activities that take due account of the size of an undertaking, the sector in which it operates, its structure and the degree of complexity of the AI system concerned (heading 8).

- take out appropriate liability insurance for their conformity assessment activities, unless liability is assumed by the Member State in which they are established in accordance with national law or that Member State is directly responsible for the conformity assessment (heading 9, slightly amended on a proposal from the Council to specify that the State concerned shall be the State of establishment of the notified body).

- be capable of carrying out all the tasks required of them under this Regulation with the highest degree of professional integrity and requisite competence in the specific field, whether these tasks are carried out by the notified bodies themselves or on their behalf and under their responsibility (section 10).

- Have sufficient internal competences to be able to effectively assess tasks performed by external parties on its behalf, for which the notified body shall have at its permanent disposal sufficient administrative, technical, legal and scientific staff with experience and expertise related to the relevant types of AI, data and data computing systems, as well as to the legally established requirements (heading 11, as proposed by the Council).

- Participate in the coordination activities referred to in Article 38 and participate in European standardisation organisations, either through direct involvement or representation, or by ensuring that they are aware of and keep up to date with the relevant standards (section 12).

- When subcontracting specific tasks connected with conformity assessment or having recourse to a subsidiary, ensure that the subcontractor or the subsidiary meets the requirements set out in Article 31 and inform the notifying authority accordingly (Article 33.1).

- Take full responsibility for the tasks performed by subcontractors or subsidiaries (Article 33.2).

- Subcontract or carry out conformity assessment activities through a subsidiary only with the agreement of the provider, and make publicly available a list of its subsidiaries (Article 33.3).

- Keep relevant documents concerning the assessment of the qualifications of the subcontractor or the subsidiary and the work carried out by them at the disposal of the notifying authority for a period of five years from the date of the termination of the subcontracting activity (Article 33(4), as proposed by the Council).

- be established under national law and have legal personality (heading 1, as proposed by the Council), without prejudice to the recognition of bodies established in third States.

- Meet the organisational, quality management, resource and process requirements necessary for the fulfilment of its tasks, as well as appropriate cybersecurity requirements (section 2). The addition of the cybersecurity requirements comes from the Council.

- Have an organisational structure, allocation of responsibilities, reporting lines and functioning that ensures confidence in the performance and results of the conformity assessment activities they carry out (section 3).

- be independent of the provider of a high-risk AI system in relation to which they perform conformity assessment activities, and of any other operator having an economic interest in the high-risk AI system under assessment, as well as of any competitor of the provider; this shall not preclude the use of assessed AI systems that are necessary for the operation of the conformity assessment body or the use of such systems for personal purposes (heading 4, final indent added on the basis of a Council proposal).

- Not be directly involved in the design, development, marketing or use of high-risk AI systems, nor represent parties engaged in such activities, or in any activity that may conflict with their independence of judgement or integrity in relation to the conformity assessment activities for which they are notified, which shall apply in particular to consultancy services (new heading 5, added on the basis of a Parliament proposal).

- be organised and operated in such a way as to safeguard the independence, objectivity and impartiality of their activities by documenting and implementing a structure and procedures to safeguard impartiality and to promote and apply the principles of impartiality in all their organisational, personnel and evaluation activities (section 6).

- Have documented procedures to ensure that their personnel, committees, subsidiaries, subcontractors and any associated bodies or personnel of external bodies respect the confidentiality of the information which, in accordance with Article 78 of the AIA, comes into their possession in the course of carrying out conformity assessment activities, except where disclosure is required by law. Therefore, it is provided that the personnel of notified bodies shall be bound to observe professional secrecy with regard to all information obtained in carrying out their tasks, except in relation to the notifying authorities of the Member State in which their activities are carried out (heading 7, the only modification being, as proposed by the Council, the addition of the reference to Article 78 of the IAR).

- Have procedures for the conduct of activities that take due account of the size of an undertaking, the sector in which it operates, its structure and the degree of complexity of the AI system concerned (heading 8).

- take out appropriate liability insurance for their conformity assessment activities, unless liability is assumed by the Member State in which they are established in accordance with national law or that Member State is directly responsible for the conformity assessment (point 9, slightly amended at the proposal of the Council to specify that the State concerned shall be the State of establishment of the notified body).

- be capable of carrying out all the tasks required of them under this Regulation with the highest degree of professional integrity and requisite competence in the specific field, whether these tasks are carried out by the notified bodies themselves or on their behalf and under their responsibility (section 10).

- Have sufficient internal competences to be able to effectively assess tasks performed by external parties on its behalf, for which the notified body shall have at its permanent disposal sufficient administrative, technical, legal and scientific staff with experience and expertise related to the relevant types of AI, data and data computing systems, as well as to the legally established requirements (heading 11, as proposed by the Council).

- Participate in the coordination activities referred to in Article 38 and participate in European standardisation organisations, either through direct involvement or representation, or by ensuring that they are aware of and keep up to date with the relevant standards (section 12).

- When subcontracting specific tasks connected with conformity assessment or having recourse to a subsidiary, ensure that the subcontractor or the subsidiary meets the requirements set out in Article 31 and inform the notifying authority accordingly (Article 33.1).

- Take full responsibility for the tasks performed by subcontractors or subsidiaries (Article 33.2).

- Subcontract or carry out conformity assessment activities through a subsidiary only with the agreement of the provider, and make publicly available a list of its subsidiaries (Article 33.3).

- Keep relevant documents concerning the assessment of the qualifications of the subcontractor or the subsidiary and the work carried out by them at the disposal of the notifying authority for a period of five years from the date of the termination of the subcontracting activity (Article 33(4), as proposed by the Council).

The new Article 32, added at the proposal of the Council, states that where a conformity assessment body demonstrates its conformity with the criteria laid down in the relevant harmonised standards or parts thereof the references of which have been published in the Official Journal of the European Union, it shall be presumed to comply with the requirements set out in Article 31 in so far as the applicable harmonised standards cover those requirements. This is a rule aimed at facilitating the accreditation of requirements by bodies, and may facilitate access to this activity on equal terms for applicants, while encouraging the adoption of standards that are expected to be of high quality.

2. Operational obligations of notified bodies. Coordination by the Commission

Article 34 of the IAR, added on the basis of a Council proposal, details the operational obligations of notified bodies.

Beyond the obvious fact that notified bodies shall verify the conformity of high-risk AI systems in accordance with the conformity assessment procedures referred to in Article 43 (point 1), it is important to note the provision that notified bodies shall carry out their activities without placing unnecessary burdens on providers and with due regard to the size of the undertaking, the sector in which it operates, its structure and the degree of complexity of the high-risk AI system concerned, while respecting the degree of rigour and level of protection required for the compliance of the high-risk AI system with the requirements of the AIA. One of the most important policy objectives has been elevated to the status of a legal standard by providing that particular attention should be paid to minimising administrative burdens and compliance costs for micro and small enterprises, as defined in Commission Recommendation 2003/361/EC (section 2).

Finally, Article 34(3) takes over the provision originally contained in Article 33, whereby notified bodies shall make available to the notifying authority and submit on request all relevant documentation, including providers' documentation, to enable the notifying authority to carry out its assessment, designation, notification and monitoring activities and to facilitate the assessment.

The coordinating role attributed to the European Commission in relation to notified bodies in Article 38 of the AIA also stands out. This is that the Commission shall ensure that, with regard to high risk AI systems, appropriate coordination and cooperation between notified bodies involved in conformity assessment procedures is put in place and properly managed, in the form of a sectoral group of notified bodies (heading 1, as proposed by the Council).

To this end, there is an obligation on the notifying authority to ensure that the bodies notified by them participate in the work of this group, either directly or through designated representatives (section 2).

And it is not surprising that, at the proposal of the Council, a new paragraph 3 has been added to Article 38, obliging the Commission to provide for the exchange of knowledge and best practices between the notifying authorities of the Member States, a provision that is certainly desirable.

3. Issue and validity of certificates

As a satisfactory outcome of the relevant conformity assessment activities, the notified body shall issue a certificate containing the contents of Annex VII of the AIA in a language which can be easily understood by the relevant authorities of the State where the body is established (Article 44(1), as proposed by the Council).

With regard to its period of validity, following the trilogues, it has been set at a maximum of five years for Annex I AI systems, and four years for Annex III AI systems; this validity may be extended for equal periods, based on additional assessments (Article 44(2), as proposed by the Council).

Of course, the outcome of the conformity assessment activities may be unsatisfactory, if the notified body considers that the AI system under assessment does not meet the legally established requirements, leading to the refusal of the body to issue the corresponding certificate, or to issue it with certain restrictions. Against such decisions, the second paragraph of Article 44(3) of the IAR, as agreed between the Parliament and the Council, ensures that an appeal procedure is available. It should be noted that both the Commission proposal and the Parliament's mandate required a legitimate interest in order to bring such a complaint, a requirement which has disappeared in the final wording.

Furthermore, where a notified body finds that an AI system no longer complies with the legally established requirements, it shall suspend or withdraw the certificate issued or impose, taking into account the principle of proportionality, unless the system provider takes appropriate corrective measures within an appropriate period set by the notified body, taking into account the principle of proportionality, a decision which it shall give reasons for (Article 44(3), as originally drafted by the Commission and endorsed by the Council).

As we have seen above, the Notified Body may be affected by changes, which will eventually affect the validity of the certificates issued.

In case of cessation of the activity of the body which issued the certificate, Article 36(3) of the AIA provides that certificates may remain valid for a temporary period of nine months after the notified body has ceased its activities, provided that another notified body has confirmed in writing that it will take over responsibility for the AI systems covered by those certificates, failing which they will cease to be valid immediately. In this case, the new notified body shall complete a full assessment of the AI systems concerned by the end of that period before issuing new certificates for those systems.

In case of suspension or limitation of a designation, the new Article 36(8) of the IAM details the circumstances under which certificates which have not been wrongly issued may remain valid, with two alternative possibilities:

- Where the notifying authority has confirmed, within one month of the suspension or restriction, that there is no risk to health, safety or fundamental rights with regard to the certificates affected by the suspension or restriction, and the notifying authority has established a timetable and the measures envisaged to remedy the suspension or restriction.

- Where the notifying authority has confirmed that no certificates relevant to the suspension will be issued, modified or reissued during the course of the suspension or limitation, and indicates whether the notified body has the capacity to continue to monitor and remain responsible for the existing certificates issued during the period of the suspension or limitation. In the event that the authority responsible for notified bodies determines that the notified body does not have the capacity to support the existing certificates issued, the provider shall communicate to the national competent authorities of the Member State in which the provider of the system covered by the certificate has its registered office, within three months of the suspension or limitation, a written confirmation that another qualified notified body temporarily takes over the functions of the notified body to monitor and remain responsible for the certificates during the period of suspension or limitation.

Finally, in case of withdrawal of a notification, the new Article 36(9) of the IAR details the circumstances under which certificates which have not been wrongly issued may remain valid for a period of nine months, with two cumulative conditions being required:

- Where the national competent authority of the Member State in which the AI system provider covered by the certificate has its registered office has confirmed that there is no risk to health, safety and fundamental rights associated with the systems concerned, and

- Another notified body has confirmed in writing that it will take immediate responsibility for these systems and that it will have completed the assessment of these systems within 12 months of withdrawal of the designation.

4. Information obligations of notified bodies

Article 45 of the IAR lays down information obligations to be fulfilled by notified bodies, both vis-à-vis notifying authorities (heading 1) and vis-à-vis other notified bodies (heading 2), following the text originally proposed by the Commission.

Heading 3 of the same article has been slightly modified during the trilogues, and the wording has been improved to refer to types of AI systems, rather than AI technologies, possibly in order to better guarantee the business secrets that notified bodies may have access to.

Finally, at the proposal of the Board, a new section 4 has been added, which subjects compliance with the reporting obligations precisely to the confidentiality regime provided for in Article 78 of the AIA, to the analysis of which reference is made.

IV. Recapitulation

In general, there is a significant similarity between the regulation of notified bodies in the AIA and the rules of the New Legislative Model, in particular Decision (EU) No 768/2008.

Instead of referring to it, it has possibly been decided to create a specific regime, although aligned with the existing one, to cover conformity assessments of high-risk AI systems, involving notified bodies, in two situations: in the case foreseen in Annex III.1 of the AIA(certain biometric processes) and in the case of notified bodies not previously designated under the harmonisation laws under the New Legislative Model (Annex I, Section A of the AIA). In the case of bodies already notified under the New Model Legislation, it would possibly have been sufficient to set out the additional requirements to be applied, as has been done in any case.

However, it is certainly striking that only in a few cases of high-risk AI systems is the intervention of a notified body actually required, leaving the rest to self-assessment by the providers, and no control at all in the remaining AI systems. It will be necessary to wait to see the results of this approach before assessing the wisdom of the co-legislators in this model of (lack of?) control.

The obligations of suppliers and deployers of high-risk systems

THE FUNDAMENTAL RIGHTS IMPACT ASSESSMENT BY DEPLOYERS OF ARTIFICIAL INTELLIGENCE SYSTEMS IN THE REGULATION

Eduard Chaveli Donet
Digital Law Specialist Attorney
Head of Consulting Strategy at Govertis, Part of Telefónica Tech

I. Introduction

If the first article of the AIA in setting out its objective states that, in addition to *"improve the functioning of the internal market"* it is to *"promote the uptake of human-centric and trustworthy Artificial Intelligence*", and at the same time emphasises that this must be *"while ensuring a high level of protection of health, safety, and fundamental rights as enshrined in the Charter"*, it stands to reason that the obligations it envisages must be aligned with these objectives. This is why the reference to fundamental rights is a constant in the recitals and in the text of the Charter.

As we know, human rights are not "something new" and over the centuries there has been an evolution in terms of their number, the beneficiaries, as well as their territorial scope; although their actual application is far from being truly global. But even where there is a "culture" and systems of human rights protection (as is the case in Europe), first industrial and then technological evolution brought opportunities and risks for human rights that had to be managed. Similarly, Artificial Intelligence (AI) is a revolution and, as such, will bring opportunities and risks that may also affect fundamental rights and have to be managed. The aim of this chapter is precisely to provide a conceptual and methodological approach to one of the obligations of the AIA to manage these risks: the Fundamental Rights Impact Assessments in certain high-risk AI systems, commonly referred to as Fundamental Rights Impact Assessments (FRIA). The purpose of FRIAs is for the implementer to identify specific risks to the rights of individuals or groups of individuals that may be affected, and to identify measures to be taken in the event that these risks materialise.

But, as we said, although the emergence of AI has occurred recently and has led to the need for its regulation, human rights have existed for a long time and there are precedents of impact assessments on human rights (HRIA), as well as social impact assessments (SIA), ethical impact assessments (EIA), as well as on some specific rights, such as the well-known example of the protection of personal data, which will serve as a basis for a methodological analysis of FRIA. There have even been methodologies and tools specifically applied to AI systems that we will also analyse before delving into the current framework contemplated by the AIA and how it has evolved in its various versions.

Furthermore, in addition to the FRIA regulated in Article 27, the AIA, in its approach to risk -as the principle that inspires it- also contemplates risk analysis as part of the AI management system (RRAAAI) in Article 9, which is dealt with in another chapter of this book, which implies intersections and possible confusions between the two, which we will try to delineate.

II. Impact assessments as a tool for weighing up fundamental rights

But before going into the adjectival part of the methodology and analysing these intersections and differences, we will start by talking about impact assessments (IAs) in general as tools for weighting fundamental rights, focusing on the fundamental rights that are the object of these assessments, and also referring to the task of weighting, which is the objective of these assessments.

From the outset, we must indicate that we will refer to both terms (human rights and fundamental rights) in an equivalent way, despite their conceptual distinction[3], in addition to the fact that – as is well known – the Spanish Constitution itself in Article 10.2 establishes that *"the norms relating to the fundamental rights and freedoms recognised by the Constitution shall be interpreted in accordance with*

[3] Human Rights have a Universal Character and Fundamental Rights, being obviously related and coinciding to a large extent, depend on how human rights are grounded in a specific field through a norm: in the European field, fundamental rights are those contemplated in the Charter and also in the constitutions. In this sense the European Union Agency for Fundamental Rights *at* https://fra.europa.eu/en/ *about-fundamental-rights/frequently-asked-questions#difference-human-fundamental-rights-access- sed 10 January 2021 ("The term "fundamental rights" is used in the European Union (EU) to express the concept of "human rights" in a specific internal EU context. Traditionally, the term "fundamental rights" is used in a constitutional setting, while the term "human rights" is used in international law. The two terms refer to a similar substance, as can be seen by comparing the content of the Charter of Fundamental Rights of the European Union with that of the European Convention on Human Rights and the European Social Charter.)*

In fact, as a curiosity, I asked a generative AI application to make a comparison between the similarities and differences in terms of rights between the UDHR and the CFREU, as well as between those included in the CFREU and those included in the Spanish Constitution, and it did so successfully, indicating the great similarities and also some differences.

the Universal Declaration of Human Rights and the international treaties and agreements on the same subjects ratified by Spain".

Having clarified the above, and in order to be able to explain what a Fundamental Rights Impact Assessment (or Human Rights Impact Assessment, HRIA) consists of, it must first be said that assessing rights means weighing them up. The weighing up of rights is carried out in the first instance by the norms themselves when they give to certain rights the status of Fundamental Rights or not. The fact that, for example, the Constitution grants different ranks to different rights is not a trivial matter because – as is well known – a series of consequences will derive from this. But this prior weighting by the Constitution is often not sufficient, especially when we talk about rights "on equal terms": for example, when we talk about weighting between fundamental rights. In each country, and on the basis of the Constitution, the legislator and, in certain exceptional cases (although increasingly frequent, the executive), when approving legal provisions, also carry out a balancing exercise. Likewise – and as Pere Simón rightly points out[4] – referring to the specific case of Administrative Law – *"if they are not resolved in the first place by the democratic legislator (Arrojo Jiménez, 2009: 27 et seq.), they may also require a balancing exercise that may proceed through the development of normative administrative activity or non-normative administrative activity – in the form of guidelines, guides, general empowerment clauses, etc.*"[5]. And obviously there is a task of weighing up which the courts carry out, and not only the ECHR and in the case of Spain the TC, but all judges and magistrates carry out this function, as can be deduced from art. 24 and 53 of the Spanish Constitution and more specifically, among others, from art. 7 of the LOPJ.

However, in addition to this task of weighing up carried out by the legislator, the courts and the administration, *ex ante* impact assessments have been imported into Europe from the Anglo-Saxon tradition, carried out by the subjects themselves who are part of the decision-making process (administrations and companies) on the means and ends that may affect the aforementioned rights. This practice stems from the fact that we live in a society in which there are an increasing number of risks[6] and an essential part of impact

[4] Simón Castellano, P., *La evaluación de impacto algorítmico en los derechos fundamentales*, Aranzadi, Cizur Menor, 2023, p. 48.

[5] As Simón Castellano, P., ob. cit. indicates (p. 48) "*such an extreme has been accepted by continental legal dogmatics and by the doctrine of the civil law tradition (Schmidt-Assmann, 2003: 2019 ff. and Franzius, 2006: 108 ff.)"*.

[6] As Mantelero, A., in ob. cit. p.13 says *"as a consequence of the transformation of modern society into a risk society, or at least a society in which many activities involve exposure to risks and which is characterised by the emergence of new risks"*.

assessments (as we will have the opportunity to discuss in greater detail later) is precisely risk management. Without prejudice to the fact that we will return to risk analysis in detail later on in the methodology, it is important to point out from the outset that risk analysis has been a common practice for years in sectors linked to "business" requirements, such as the financial and insurance sectors. It is also a subject that has historically been carried out on the basis of standards, both general (such as ISO 31000:2018, on risk management) and other ISO standards that have brought risk management to various IT fields, such as ISO/IEC 27001:2022 on information security, ISO 22301:2020 on business continuity; or in data protection, ISO 27701:20021 on privacy management systems, or ISO/IEC 27018:2014 on the processing of personal data in the cloud, among others. All of them refer to risk management. Likewise, risk analysis has already begun to land in legal provisions, such as the National Security Scheme (NSS)[7] and is proliferating in other areas such as criminal compliance, money laundering, data protection and now in relation to Artificial Intelligence. Indeed, the AIA coins a risk-based approach – according to its Recital 26 – *"to introduce a proportionate and effective set of binding rules for AI systems"* by tailoring *"the type and content of such rules to the intensity and scope of the risks that AI systems can generate"*. This is done by the AIA through the following formula that some authors have criticised[8] : "*to prohibit certain unacceptable AI practices, to establish requirements for high-risk AI systems and obligations for the relevant operators, and to lay down transparency obligations for certain AI systems"*.

1. Differences between risk analysis and impact assessments. The example of data protection

In order to make some necessary conceptual clarifications on the concept of risk analysis (RRAA) and impact assessments (IA), we will use the example of data protection (with obvious links to information security), given that it is perhaps the area that has reached the greatest maturity in our context in relation to the generalisation of risk analysis and impact assessments (*ex art.* 35 of the GDPR), given that it is a matter that "applies" to all sectors of activity. This seems important to us because: on the one hand, many professionals

[7] The National Security Scheme in its initial version of RD 3/2010 modified by RD 951/2015 already contemplated it, and obviously the new NSS approved by Royal Decree 311/2022, of 3 May, has maintained it.

[8] Vid. Mantelero, A., in ob. cit. p. 173 which indicated *"Although this is effective in terms of political impact and acceptability, it is a weak form of risk prevention. The Proposal makes a rather rigid distinction between high-level risk and the rest, providing no methodology for assessing the former, and largely exempting the latter from any mitigation (with the limited exception of transparency obligations in certain cases)"*.

who are now approaching AI come from data protection backgrounds, and also because AI systems can and often do involve the processing of personal data. However, despite such a comparative approach, as we will see in the next section, precisely in the case of RRAAAI and FRIA these considerations cannot be *applied mutatis mutandis*.

Impact assessments and risk analyses are two different, but intimately linked, things, as the AEPD (Spanish Data Protection Agency) says (and is developed in detail in the Guide[9]):

"Risk analysis and risk management are procedures that enable organisations to identify and anticipate potential adverse or unintended effects of processing on the rights and freedoms of data subjects. This management should enable the responsible person to take the necessary decisions and actions to ensure that the processing complies with the requirements of the GDPR and the LOPDGDD, guaranteeing and being able to demonstrate the protection of the rights of data subjects.

The GDPR states that where a type of processing is likely to involve a high risk, the responsible person must carry out an impact assessment, a process that allows organisations to identify the risks that a system, product or service may pose to the rights and freedoms of individuals and, having carried out that analysis, to address and manage those risks before they materialise.

Risk management and DPIA (Data Protection Impact Assessment) are closely linked processes, the latter being a specificity within the former. Thus, DPIA cannot exist without being part of risk management, so while risk management is mandatory for all processing, the specific obligations set out for DPIA are exclusively for high-risk processing".

Likewise, data protection is clearly related to information security, insofar as personal data are an information asset and must be protected with the appropriate security measures. For this reason, and to visualise it with an example in the public sector: the NSS, to which we have referred, obliges organisations to carry out a risk analysis of the information within its scope, in order to determine the measures to be applied on the basis of taking into account the possible impact on it of five dimensions of security: confidentiality, integrity, authenticity, availability, and traceability.

In this case, for example, risk analysis helps to protect information security and does not take into account the specific risk that exists for the processing of personal data. This is why the NSS itself determines that "*where a system involves personal data, the controller or processor, on the advice of the data protection*

[9] https://www.aepd.es/prensa-y-comunicacion/notas-de-prensa/aepd-publica-nueva-guia-gestionar-riesgos-y-evaluciones-impacto

officer, shall carry out a risk analysis in accordance with Article 24 of the GDPR and, in the cases set out in Article 35, a data protection impact assessment".

Therefore, and in conclusion: if we are talking about personal data protection, the risk analysis to be carried out for all personal data processing can be based on the one that has been carried out with another framework (in this case the NSS, just as another ISMS such as the one based on ISO 27001 can be taken into account, but the additional protection risks that in certain cases and as a result of the processing of certain types of data (for example special categories of data) must also be taken into account.

But: Can we apply this GDPR distinction to the case of the AIA? Let's look at it below.

2. Differences between the risks to be addressed in the risk management system of Article 9 vs. the fundamental rights impact assessment of Article 27 AI Act

Being the above the theoretical relationship between the DPRRAA (Data Protection Risks Analysis) and the DPIA (Data Protection Impact Assessment) that the AEPD has collected in its guide[10] (echoing articles of the GDPR and also of the WP of the EDPB), it also describes not only their relationship but also the peculiarities of the DPIA that differentiate it from the RRAA in data protection. I would like to thank especially Jordi Morera and María Loza who in a choral work have helped me to work on a position of "common understanding".

Therefore, and only for the purpose of this distinction between DPRRAA and DPIA to inspire the comparative process between RRAAAI required by Article 9 of the AIA and FRIA required by Article 27 of the AIA, we will follow (for scripting purposes only) the structure proposed by the AEPD for DPIA:

On the one hand, the DPIAs are enforceable *"when there is a high risk to rights and freedoms"*. In the case of RRAAAIs and FRIAs, the essential difference is that both start from a situation of "High Risk" (since a High Risk AI is being used), but FRIAs are also only required of the implementers of certain processes that, because they are public or because of the subject matter, which we will see later, are understood to pose a "higher" risk, requiring additional supervision by the authorities (which is why they are included in documentation and in the process of review by the Authorities). This is with-

[10] https://www.aepd.es/documento/gestion-riesgo-y-evaluacion-impacto-en-tratamientos-datos-personales.pdf

out prejudice to the fact that providers of high-risk systems who are subject to the RRAAAI of Article 9 must register the system, so that oversight applies to all high-risk providers (Art. 51). Even providers that do not consider themselves to be high risk are also required to register.

At the level of the parties obliged to carry it out, and as we will develop later, just as the DPIA is a specific obligation of the controller, without prejudice to the assistance of the processor, to which we will return later; in the case of the RRAAAI it is an obligation of the providers in its broad concept contemplated by the AIA (*see* Recital 96), although it may also be carried out by those who carry out the deployment or users of the AI system. But FRIA is an obligation of the deployer, notwithstanding the fact that it may be based on the provider's RRAAAI.

In the case of the DPIA, an analysis of the necessity and proportionality of the processing in relation to its purposes is required. The RRAAAI does not require this analysis of necessity and proportionality; and although the text of the AIA for FRIA does not require it either, as we shall see, it is required by certain methodologies that have already existed as precedents and, in our case and as we shall examine in more detail later, we believe that it should be carried out, at least in a *soft* form.

In the case of the DPIA, this is required prior to the start of processing activities, which is not required in the DPRRAA. In the case of RRAAAIs, nothing is said, and although it seems logical that this should be done before any high-risk AI system is put in place, it is not required, unlike in FRIAs where it is expressly stated; and perhaps this is another difference.

Unlike the DPRRAA, the DPIA requires the advice of the DPO, if appointed. In contrast, the RRAAAI does not require the advice *per se* of any role, without prejudice to the appropriateness of the intervention of various roles to which we will refer later. In the case of FRIAs, they will require the advice of the DPO when there is a processing of personal data that requires an DPIA.

In the case of the DPRRAA, nothing is said in this respect, and in the case of the DPIA, the opinion of the interested parties, or their representatives, where appropriate, must be sought in the risk management process, justifying, where appropriate, the inappropriateness or limitation in the communication of information. For its part, in the case of RRAAAI, Recital 96a of the AIA indicates that, when identifying the most appropriate risk management measures, the provider must document and explain the decisions taken and, where appropriate, involve external experts and interested parties. In the case of FRIA, Recital 96 states that "where appropriate, in order to gather the relevant information necessary to carry out the impact assessment, implementers of high-risk AI systems, in particular where AI systems are used in the public sector, may involve relevant stakeholders, including representatives of groups of persons likely to be affected by the AI system, independent experts, and civil society organisations in carrying out such impact assessments and in designing the measures to be taken in the event of materialisation of the risks".

In the case of the DPIA, unlike the DPRRAA, its outcome should be taken into account to assess the feasibility or unfeasibility of the processing from a data protection point of view. In the case of RRAAAI and DPIAIA it seems logical that in both cases it should be taken into account to analyse whether the AI system remains operational or not, as one might think in the case of data protection; but in the case of FRIA, unlike RRAAAI, it is made clear that it should be carried out before the first use.

Unlike the DPRRAA, in the case of the DPIA, depending on the level of residual risk, the data controller is obliged to carry out a Prior Consultation (art. 36 GDPR) with the supervisory authority. In the case of RRAAAIs nothing is said, but in the case of FRIAs (not linked to residual risk but in any case) the market surveillance authority must be notified of the results of the assessment, by submitting the completed template referred to in Article 27.5, except in certain cases, which we will refer to later. This is an important point, if yes or yes, I have to notify the result of the FRIA, it could be interpreted that the FRIA joins both certain features of the DPIA and the prior consultation of Art.36, since, by registering the system with the Assessment, the supervisory authorities (based on the general powers they have, e.g., Art. 67), could order the adoption of corrective measures on the AI to further mitigate the risks.

Another difference between the DPRRAA and privacy DPIA is that in privacy (in both cases) it is necessary to take into account the specific situations of the processing (this can be read repeatedly in the AEPD guide and is deduced from the GDPR). In other words, the type of analysis. In the case of the RRAAAIs, it refers to "known and predictable risks", both from "normal" use and misuse. And taking into account that the obliged is the solution provider, but not the implementer/user of the solution, it will be a "general" risk analysis, i.e., the typology of threats and taxonomy of threats will be on the product itself and its intended uses and intended misuses, but will not be grounded to the specific use case of a given company. The analysis is more "customised" since it already involves the deployer/user, and importantly the legislator indicates: "*shall perform an assessment of the impact on fundamental rights that the use of the system may produce. For that purpose, deployers shall perform an assessment consisting of... a description of the deployer's processes in which the high-risk AI system will be used in line with its intended purpose*". So the FRIA is a much more use-case focused analysis. As I have indicated in other comments, the emphasis is much more on describing the processes and the concrete use (groups affected, frequency, etc.).

In other words, although the reference to the DPRRAA and the DPIA has served as a comparison, precisely this comparison and the reading not only of Articles 9 and 27 of the AIA but also the general reading of the AIA leads us to understand that the comparison between DPRRAA and DPIA (Articles 24 and 32 of the GDPR in relation to Article 35 of the GDPR) is not exactly applicable to the RRAAAI and FRIA. Whereas the DPIA of 35 GDPR are clearly an additional requirement with respect to the general requirement (for both public and private sector) with respect to DPRRAA; on the other hand, in the case of the AIA, it seems to imply that Article 9 AIA intends that all High Risk AI solutions have documented risks managed by the provider of the solution (as occurs in other products) but evidently not tailored to the specific use case of a particular company that deploys them and according to its particularities, but at the level of what is reasonably foreseeable and that it is updated according to market monitoring. On the other hand, Art. 27, when referring to FRIAs, does seem to be oriented towards analysing the risks in the specific case of use, but only applicable to the case of use that involves or is connected to the exercise of public functions and specific cases that are indicated below.

To bring it down to an example, assuming that provider A develops a Chatbot (we assume that it is high risk), it will have a risk analysis of different scenarios of use, but if this Chatbot is involved by Administration A, it will have to make a FRIA on its use case and Administration B also (for example, the first wants to implement it on citizens and the second on employees, it would vary the circumstance of Art. 27.a 1. (c) (groups of interested parties), the result of this of two FRIA by two different implementers/users could be different (although the legislator already hints at the possibility of applying analogy to similar cases already validated).

In short, the FRIAs of Art. 27 act as a safeguard to avoid abuses in certain cases, but also as a mechanism for legal certainty and to promote the use of AI in the sense that, once a use case has been validated, it can already be taken into account to validate FRIAs on similar cases.

3. Typologies of impact assessments

Having established the intimate relationship between RRAA and impact assessments, it is necessary to understand the different models of impact assessment that exist before focusing on the FRIAs proposed by the AIA.

Mantelero[11] refers to the closest examples that serve as an approximation: HRESIA or HRIA (Human Rights, Ethical and Social Impact Assessment), PIA (Privacy Impact Assessment) or DPIA (Data Privacy Impact Assessment), SIA (Social Impact Assessment) and EtIA (Ethical Impact Assessment); he analyses the characteristics, similarities and differences of each model, their advantages and possible disadvantages that they entail. However, we will focus on HRIAs. As fundamental rights have existed for years and prior to the existence of AI and its recent emergence, they have been and are subject to risks in environments other than AI. For this reason, there have been impact assessments on different sensitive issues: for example, impact assessments in the environment were already considered in the 1960s, and are currently a legal requirement in many countries. In the case of Spain, for example, they are included in Law 21/2013, of 9 December, on environmental assessment. Beyond the debate on the need for environmental protection to be a human right, the fact is that it is included in the Charter of Fundamental Rights of the European Union (CFREU) and there is also an evident concern for the environment that has been increased by climate change and that has also been projected in the AIA itself in various recitals[12] and articles. It is no coincidence that ISO 42001:2023, to which we will refer later on about AI, also makes reference to the environment and climate change. However, although we will go into more detail later, unlike other rights, in the case of the environment it is an impact assessment on more objectifiable aspects (water, soil, air, etc.).

Historically, HRIAs have been more commonly carried out in relation to certain activities that are "sensitive" to certain rights, such as open-pit mining, oil or gas operations, factories and other activities where – not infrequently – such activities have taken place in countries with a significant human rights deficit[13] and carried out by the companies that exploit such activities

[11] Mantelero, A., *Beyond Data: Human Rights, Ethical and Social Impact Assessment in AI*, Information Technology and Law Series, 2022.

[12] For example, the AIA provides in Recital 27 *that "Social and environmental welfare" means that AI systems are developed and used in a sustainable and environmentally friendly manner as well as in a way to benefit all human beings, while monitoring and assessing the long-term impacts on the individual, society and democracy. The application of those principles should be translated, when possible, in the design and use of AI models".*

[13] One aspect that the International Business Leaders Forum (IBLF) and the International Finance Corporation (IFC), in partnership with the UN Global Compact Office, focus on is the territorial level:

- An area with weak governance.
- A country in a precarious state and/or affected by conflict.
- An area where human rights commitments are poorly implemented.

and which belong to more developed countries. In addition, and not infrequently, accusations have been made that these operations are an attempt at image-cleansing in the face of news reports of allegations.

Some examples of these sectors can even be found in well-known Spanish multinationals, such as Iberdrola or Repsol (although with a different scope: general in the first case and affecting a specific operation in the second one) and also with different levels of detail in the information published, as can be seen. In many cases the approved policies and HRIAs have been based on the UN Guiding Principles. There have obviously been HRIAs in other private sectors as well, but historically less common.

Another common example of HRIAs are those carried out by certain NGOs (such as Oxfam or the Red Cross, for example), which are usually published but which have not been free from criticism of possible bias and accusations of partiality.

The public sector is no stranger to HRIAs either, with proliferating cases such as the one led by the Secretariat for Relations with the Courts on the General State Administration[14].

In other cases, approaches have focused more on the affected subjects, such as children[15] or other vulnerable groups, rather than the specific sectors of activity. And this has also been projected onto AI. It is no coincidence that the AIA mentions precisely children, for example in Recital 48[16] and that Article 9 risk management[17] is mentioned.

We are no strangers to the tensions arising from different views on human rights and the debate on universalism and cultural relativism in human rights, but human rights, even with the usual differences, provide a broadly

- An area with high environmental and/or social risks and impacts.

- An area inhabited by vulnerable local communities (e.g., indigenous peoples).

[14] https://www.abogacia.es/wp-content/uploads/2013/01/INFORME-DE-EVALUACION-DEL-PLAN-DE-DDHH.pdf

[15] Such as in the case of children, where in 2012 UNICEF adopted guidance on integrating children's rights into impact assessments and other evaluations.

[16] Recital 48 of the AIA refers to children having specific rights enshrined in Article 24 of the EU Charter and in the UN Convention on the Rights of the Child (elaborated in UNCRC General Comment 25 with regard to the digital environment), both of which require that children's vulnerabilities be taken into account and that they be provided with the protection and care necessary for their well-being.

[17] Article 9.8. When implementing the risk management system described in paragraphs 1 to 6, providers shall take into account whether, in view of its intended purpose, the high-risk AI system may adversely affect children under the age of 18 and, where appropriate, other vulnerable groups of persons.

acceptable, and applicable, framework for reference in relation to AI impact assessments.

In addition to HRIAs, there have also been examples of impact assessments that integrate human rights with social and ethical aspects, for it is not for nothing that ethical and social values are fundamental to grounding human rights in each geographical and cultural context.

III. Background on fundamental rights assessments of AI systems prior to the AI Act

As we have seen, impact assessments on fundamental rights, as well as social and ethical impact assessments already existed prior to the emergence of AI, but before the approval of the AIA, it was necessary to formulate conceptual and methodological approaches, both doctrinal and practical, on Algorithmic Impact Assessments.

I think it is fair to highlight two of the few, but magnificent publications on the subject that have been published and which are cited repeatedly in this chapter: Mantelero's[18] and Pere Simón's[19], without prejudice to the fact that there are more, and we will comment on some of them.

In the case of Pere Simón, he classifies the possible methodologies according to different criteria:

- Depending on the consequences: The rights and legitimate interests potentially affected.
- Depending on the technology used.
- Depending on the involvement of third parties: providers, customers, sellers, auditors.
- And according to the main risks, which are grouped into three blocks: risk of misinformation, risk of discrimination (*disparate impact assessment*), risk of defencelessness and second chance.

For his part, Mantelero, in the aforementioned work, makes a concrete commitment to HRESIA (Human Rights, Ethical and Social Impact Assessment), a hybrid model that takes into account *"both the ethical and social impact of a technology together with the legal and human rights dimensions*" and that allows "*combining the universality of human rights with the local dimension of social values"*. It argues that traditional HRIA reports often describe the risks encountered and their potential impact, but without a quantitative assessment, and of-

[18] Mantelero, A., ob. cit.

[19] Ob. Cit. Simón Castellano, P.

fer recommendations without ranking the level of impact, leaving it to decision-makers to define an appropriate action plan.

In this view, he continues, *"ethical and social values are viewed through the lens of human rights and serve to go beyond the limitations of legal theory or practical application to effectively address the most pressing issues relating to the social impact of AI"*. Indeed, such ethical and social values are taken into account when interpreting human rights by authorities and courts.[101]

This is – according to Mantelero – not a technological assessment, but an assessment based on different rights and values, which *"can respond to the demand for a broader protection of individuals in the context of AI and better respond to the growing demand for AI"* and *"is consistent with studies in the field of collective data protection that point to the importance of these non-legal dimensions in the context of data-intensive applications",* which is consistent with the fact that AI systems often take into account data that affect groups or collectives.

In addition to the doctrinal approaches in Spain, it should be borne in mind that various countries and institutions have developed methodologies and tools, which Pere Simón discusses in the aforementioned book:

- In Canada, the Risk Impact Assessment Tool (RIAT).
- The US Government's AIA tool.
- In the Netherlands the model FRAIA [20].
- And others such as the IEOAC's PIO (Principles, Indicators and Observables) model, or the Ada Lovelace Institute's proposals and the European Law Institute's Model Rules.

It is not our intention here to go into each of these tools in depth, but we do make some basic considerations about some of them, including the appropriate references for further information and subsequently some of their characteristics are taken into account when analysing the proposed model.

With regard to the Canadian RIAT, not only is the tool available *online*[21] but also the explanation of the model[22]; and in addition to other aspects that will be indicated later in relation to the methodology, I would highlight the areas that are the object of analysis:

- The rights of individuals or communities.
- The health of individuals or communities.

[20] https://www.government.nl/binaries/government/documenten/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms/Fundamental+Rights+and+Algorithms+Impact+Assessment.pdf

[21] https://open.canada.ca/aia-eia-js/?lang=en last consulted on 12/03/2024.

[22] https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html last consulted on 12/03/2024.

- The economic interests of individuals, entities, or communities.

- And the continued sustainability of the ecosystem.

Without denying the value of being a pioneering tool and the simplicity of its model, the truth is that when entering into it, some aspects and rights that have been indicated are analysed, but there are some rights that are not analysed and – likewise – the depth of the analysis of the risks of each right seems to me to be too simple, which may detract from its validity[23].

For its part, the US is also a country of absolute reference in AI, without prejudice to the differences it has with Europe, which is going to "arrive later than Europe" to have a specific and complete regulation of AI. There have been different initiatives, the most notorious being the recent Executive Order signed by Joe Biden on 30 October 2023, which some have described as "*a time of war law"*, and which, together with other pre-existing ones, constitutes a germ of what will probably be the future regulation in the US. Among these pre-existing initiatives we can cite the *Algoritmic Impact Assessment*, an online tool to help companies manage AI risks.

On the other hand, the aforementioned FRAIA (*Fundamental Rights and Algorithm Impact Assessment*) model from the Netherlands is a manual made available by the Ministry of the Interior and Relations to assist organisations in making decisions on the use of AI systems. FRAIA distinguishes the following phases:

– Part 1. Why? Here the "Why" of the intention of the algorithm is analysed. What are the motives and effects and the underlying values.

-What? (input) Here the focus is on the form of what, the object, the algorithm. This part is divided into two subparts:

1. Part 2A concerns the input to the algorithm: the data to be used and the corresponding preconditions.

2. Part 2B concerns the algorithm itself. For example: what kind of algorithm is used and what are the preconditions for a responsible use of the algorithm.

– Part 3. How? This part refers to how implementation, use, monitoring, and results take place.

[23] Obviously, these are different legal provisions and a different tool, but we would like to point out here that the rights that are the object of protection are also different. For example, there are rights recognised in the ECHR or in the UDHR that the RIAT directly or indirectly evaluates: life, human dignity, psychological and mental integrity, the right to work and fair working conditions, the protection of personal data or the right to health and medical care; but there are others that the RIAT does not directly or indirectly take into account, such as the right to education, intellectual property, the right to housing or freedom of expression and information.

– Part 4. Fundamental rights Roadmap. This incorporates a "fundamental rights roadmap" with a two-fold objective:

1. Serves as a tool to identify whether the algorithm to be used will affect fundamental rights;

2. If so, facilitate a structured discussion on whether there are opportunities to prevent or mitigate this interference with the exercise of fundamental rights, and whether this would be considered acceptable.

There are other examples of methodologies, such as "*Deepfakes, Phrenology, Surveillance, and More ¡ A Taxonomy of AI Privacy Risks*"[24], which is an approach to AI events integrated with the Taxonomy of Privacy events; or others such as PLOT4ai,[25] which is a tool that contains a collection of 86 different threats and a threat modelling methodology, being able to select different or all catalogues (Technique & Processes, Accessibility, Identifiability & Linkability, Security, Safety, Unawareness, Non-compliance and Ethics & Human Rights). For example, for this part of Ethics & Human Rights, 17 questions are asked. Questions are asked and based on thes, the corresponding threats are determined, and all of this is reflected in a report.

Likewise, and before getting to the core object of this chapter (FRIA in AIA), it is necessary to refer to the ISO norms which, as standards, have not only standardised aspects related to what we have referred to as RRAA methodology, for example, but also specifically in terms of AI. There are several ISOs that refer to AI and that should be taken into account when approaching the methodology of EEIIDDFFIA:

- ISO/IEC 42001: 2023 *Information technology Artificial Intelligence* Management system, which specifies the requirements for establishing, implementing, maintaining and continually improving an Artificial Intelligence management system (IAMS) in organisations, of which I think it is worth highlighting ANNEX D, which refers to something that will become commonplace, such as integration with other ISO standards such as 27001, 27701 or 9001.

- ISO/IEC TR 24030:2021 *Information technology Artificial Intelligence* (AI) which provides a collection of use cases for Artificial Intelligence (AI) in various domains will be replaced by ISO/IEC TR 24030, which is in the process of being published.

- ISO/IEC 22989:2022 Information technology -Artificial Intelligence-Artificial Intelligence concepts and terminology, which establishes terminology and describes AI concepts.

- ISO/IEC 23894:2023*, guidance on risk management in AI,* which provides

[24] https://arxiv.org/pdf/2310.07879.pdf

[25] https://plot4.ai

guidance on managing risks related to Artificial Intelligence (AI) in organisations.

- And the recent ISO/IEC TR 5469:2024 *Artificial Intelligence Functional safety and AI systems* describes the properties, risk factors, available methods and processes related to the use of AI systems to design and develop safety-related functions.

- UNE CEN/CLC ISO/IEC/TR 24027:2023, which addresses biases in relation to AI systems, should also be taken into account.

All of them can: On the one hand, help to have muscle around certain concepts with respect to AI; but – above all – to rely on risk management methodology and management system standards that have a global vision and scope and can complement the AIA, both for the RRAAAI referred to in article 9 AIA and for the FRIAs that are the subject of this chapter.

Furthermore, ISO/IEC 42002:2023 provides us with a concept of AI System Impact Assessment as a *"formal, documented process by which an organisation developing, providing or using products or services that use Artificial Intelligence identifies, evaluates, and addresses the impact on individuals, groups of individuals, or both, and societies"*.

IV. The Fundamental Rights Impact Assessment of High-Risk Artificial Intelligence Systems in the Act

1. Development, processing and final content of the articles of the Act involved

FRIAs have been a "recent" introduction in the legislative process since they first appeared in the Parliament's version. The first change in the final version is that, unlike the Parliament's version, which required those responsible for the deployment of high-risk systems to carry out FRIAs, the final version restricts them to specific deployers. These include bodies governed by public law, private operators providing public services, and operators deploying high-risk systems, as mentioned in Annex III, points 5(b) and (d) (to which we will refer later). We believe this restriction is inappropriate, as it excludes numerous cases in the private sector that, in our opinion, should be covered by this obligation:

- As for the content of the evaluation, the structure remains essentially the same, although some adjustments have been made in the final version:

- On the one hand, in the description, in addition to the intended purpose of the AI system, there is a need to refer to the implementer's processes in which the system will be used. This is a logical provision because an understanding of the process followed in an AI system is essential when it comes to assessing risks and measures. Also, the time period in which it will be used is added to the frequency and the reference to the geographical scope is removed. The reference to geographical scope may be irrelevant in the context of assessments under the AIA because, whatever its scope, the rights that apply in the analysis are the fundamental rights under the EU CFDD, which is what the AIA regulates. However, if a FRIA of a broader scope, both geographically, or if ethical and social aspects are to be added beyond the minimum scope of the AIA, reference to and knowledge of the geographical scope and therefore of that context will be necessary.

- On the other hand, as regards the periodicity in which the FRIA should be carried out, the final version does not differ much from the Parliament's version as both agree that "*it shall apply to the first use of the high-risk AI system and that the implementer may, in similar cases, rely on previously conducted fundamental rights impact assessments or on existing impact assessments carried out by the provider*". They also agree that if, during the use of the high-risk AI system, the implementer considers that the criteria listed in paragraph 1 are no longer met and that they led to the need to carry it out, further action will have to be taken, but unlike the Parliament's version which said that *"a new fundamental rights impact assessment shall be carried out"*, the final version says that the implementer *"shall take the necessary steps to update the information"*.

- The part on risks is simplified by moving the reference to fundamental rights as the object of analysis to the beginning of the article, and thereby also placing them – as appropriate – at the epicentre of the analysis. It also modifies and clarifies that these risks can be to the categories of individuals and groups that may be affected by their use in the specific context and not only, as mentioned in the Parliament's version, to "*marginalised or vulnerable groups*", which of course may be a subset of those. Finally, it adds at this point, rightly in our view and as will be referred to later, that in the impact assessment the deployer must take into account the information provided by the provider under Article 13.

- In the final version, the reference in the Parliament's version to *"the reasonably foreseeable adverse impact of the use of the system on the environment"* is deleted. This is an inclusion that – as we have also noted in this chapter – appears in various parts of the AIA as a reinforcement, although protection of the environment is itself a fundamental right under the CFREU and is therefore not necessary.

-The final point to note in relation to content is that the Parliament's version referred to the need to include a detailed plan on how the harm and negative impact on the identified fundamental rights will be mitigated. The final text no longer includes this reference, but the risk analysis of an impact assessment suggests the need for such an action plan. The only mention of mitigation as a risk management measure is reductionist, as we will discuss later. While mitigation is a common approach to risk management, it is not the only one.

- Both texts expressly envisage, among the measures to be adopted, both the system of governance and human supervision. However, the final version:

a) On one hand, it has added the reference to complaint mechanisms, which I understand to be a legal obligation, perhaps an extension of the subjective scope of the Directive and the State laws that implement it, not only because of the matters subject to complaint but also because of the AI system used, but I do not understand its place in an impact assessment, beyond reinforcing (or establishing its requirement, as the case may be).

b) On the other hand, it has removed the reference in the Parliament's version to complaints handling and redress, which seems logical to me, as these are aspects covered by the AIA and referred to in other chapters of this work, and it makes no sense for them to be contained in an impact assessment.

- On the other hand, the final version has removed the reference in the Parliament's version that, "if a detailed plan to mitigate the identified risks cannot be identified in the course of the assessment, the implementer shall refrain from putting the high-risk AI system into use and shall inform the provider and the national supervisory authority thereof without undue delay". In this regard:

Firstly, the reference to refraining from putting it into use is actually a logical provision, but it may be redundant because it is obvious that if there is no risk treatment plan and, furthermore, it could be added, the risk threshold cannot be lowered to an acceptable risk that has been defined, the system cannot be used.

B) Secondly, a distinction must be made:

- With regard to the existing provision for communication to the authority in these cases, which has been eliminated, we understand that the purpose was the same as that contemplated in the GDPR with prior consultations of the DPIA[26], although I am afraid that the experience of the GDPR on this

[26] As Recital 94 of the GDPR notes "*Where a data protection impact assessment indicates that the processing would, in the absence of safeguards, security measures and mechanisms to mitigate the risk, result in a high risk to the rights and freedoms of natural persons and the controller is of the opinion that the risk cannot be mitigated by reasonable means in terms of available technologies and costs of implementation, the supervisory authority should be consulted prior to the start of processing activities*" and accordingly Article

point – I believe – has not confirmed its practical usefulness, at least in Spain (review).

- And as for the communication of such assumptions to the provider, the only sense I understand it could have is because it has been warned that it is a risk derived, not from the deployment, but from the product provided by the provider. And perhaps that is also why the Parliament's version of this article states that *"national supervisory authorities shall, in accordance with Articles 65 and 67, take this information into account when investigating systems that present a risk at national level*[27] ". And perhaps for this reason, when referring to providers and not implementers, and given that for them such consultation is maintained in the aforementioned articles, it has been removed from this article that affects implementers.

The Parliament's version also included a now deleted obligation for FRIAs (with the exception of SMEs, which it indicated could do so voluntarily, as well as in some cases concerning public authorities) for the implementer to "*notify the national supervisory authority and relevant interested parties" and "involve representatives of persons or groups of persons likely to be affected by the AI system"* and mentioned some examples such as: *"equality bodies, consumer protection bodies, social partners, and data protection bodies, with a view to receiving input for the impact assessment".* Also, as mentioned elsewhere in this chapter, stakeholders should be involved in the FRIA, but there is no obligation to publish or communicate the results to them, which, as we will see later, can be good practice.

-Also the Parliament's version required that where the deployer was a public authority or an undertaking referred to in Article 51(1a)(b) ("implementers which are undertakings designated as gatekeepers under Regulation (EU) 2022/192"), to publish a summary of the results of the impact assessment. This obligation has also disappeared and as with the publication and communication to stakeholders of the outcome, and as mentioned later and – omitting parts that may be sensitive or in summary mode – may be a good practice as well. In contrast, the final version has added a further novelty that

36.1. GDPR provides that *"The controller shall consult the supervisory authority prior to processing where a data protection impact assessment under Article 35 indicates that the processing would result in a high risk in the absence of measures taken by the controller to mitigate the risk" in order "to be able to advise the controller".*

[27] Article 67.1 of the AIA refers to where, having carried out an assessment in accordance with Article 65, after consulting the relevant national public authority referred to in Article 64.3, the authority finds that, although a high-risk AI system complies with this Regulation, it presents a risk to the health or safety of persons, fundamental rights or other aspects of public interest protection, it shall require the relevant operator to take all appropriate measures to ensure that the AI system concerned, when placed on the market or put into service, no longer presents that risk without undue delay, within such period as it may specify.

was not foreseen in the Parliament's version: The obligation for the *deployer to "notify the market surveillance authority of the results of the assessment, submitting the completed template referred to in paragraph 5 as part of the notification",* with some exceptions. In the case referred to in Article 47.1 (which concerns AI systems relating to public safety or the protection of human life and health, the protection of the environment and the protection of key industrial and infrastructure assets) and in line with this novelty has added a paragraph stating that *"the AI Office shall develop a model questionnaire, including by means of an automated tool, to facilitate users' compliance with the obligations of this Article in a simplified manner".*

- Finally, as regards the relationship with the DPIAs where the AI system involves the processing of personal data, the Parliament's version indicated that the deployer would carry out the DPIA-PDFAI in conjunction with the data protection impact assessment and that *"the data protection impact assessment would be published as an addendum*" to it. The final version has retained the fact that they are to be carried out jointly but, not insignificantly, makes several significant changes:

- On the one hand, part of the assumption for the assessment to be joint is *if* any of the obligations established for FRIA are already being met by the DPIA, because we understand that it is assuming the possible relationship between AI systems and personal data processing and the possible need to carry out a DPIA but also, as we will go into in more detail later, that – given the development and musculation of the DPIA- it is possible that joint means obviously coordinated but not necessarily together.

- And, perhaps also for this reason, it removes the reference to the DPIA being published as an addendum to the FRIA, both because it is possible that, being coordinated, it would be a separate report and because the obligation to publish FRIAs has been removed from the final text and there is no obligation under data protection law to publish PIAs, notwithstanding that it may be considered good practice to publish parts of it or extracts, managing the risks both to the security of the information, the AI system and other legitimate rights and interests of the organisation, such as trade secrets, for example.

2. Analysis of FRIA in the Act

2.1. Subjective scope: Who is obliged to carry it out and who is involved in it?

As discussed throughout this book, the AIA considers different operators in the chain of an AI system (the provider, the product manufacturer, the deployer, the authorised representative, the importer or the distributor) and with different obligations.

In the case of impact assessments, the obligation to carry out an impact assessment is incumbent on the deployer. The AIA clarifies that this obligation applies to certain specific deployers: "bodies governed by public law or private operators providing public services and operators deploying high-risk systems referred to in Annex III(5)(b) and (ca)").

Therefore:

1. On the one hand, it must be carried out by public law bodies (it is important to take into account Laws 39 and 40/2015) for all high-risk AI systems.

2. On the other hand, private operators providing public services (again, it is important to take into account Laws 39 and 40/2015) with respect to all AI systems that refer to these public services. In fact, Recital 96 gives some examples[28] but which cannot be understood as a *numerus clausus.*

And on the other hand (irrespective of the public or private nature of these entities) and by reason of the purpose of the systems, certain operators deploying high-risk systems referred to in Annex III(5)(b) and (ca) and which (consistent with Recital 96) are:

1. On the one hand, *"AI systems intended to be used to evaluate the creditworthiness of natural persons or establish their credit score, with the exception of AI systems used for the purpose of detecting financial fraud".* Recital 96 gives as examples "banking or insurance institutions".

2. and on the other hand "*AI systems intended to evaluate and classify emergency calls by natural persons or to be used to dispatch, or to establish priority in the dispatching of, emergency first response services, including by police, firefighters and medical aid, as well as of emergency healthcare patient triage systems".*

Likewise, Article 27.2 as Recital 86 provide that *"in similar cases the deployer may rely on previously conducted fundamental rights impact assessments or existing impact assessments carried out by provider",* so it is a clear allusion that also the providers of such AI systems mentioned in the previous scope should carry it out, or perhaps there is a terminological confusion and they should rely on the RRAAAIs that, ex. Article 9 AIA, must be carried out by the providers. In any case, it seems clear that the providers must assist the implementer, as is the case with data protection between data controllers and data processors[29].

[28] *Recital 96 adds "Services important for individuals that are of public nature may also be provided by private entities. Private entities providing such public services are linked to tasks in the public interest such as in the areas of education, healthcare, social services, housing, administration of justice".*

[29] In the same way as in data protection, where the AEPD GUIDE on RRAA and DPIA cited above, and after stating that the obligation to carry it out lies with the controller, also mentions the obligation of processors to assist the controller (in line with the provisions of the GDPR (Recital 95) *"The processor should assist the controller, where necessary and upon request, in ensuring*

2.2. Team to be involved

When thinking about the team that should address it, it is essential to involve all relevant actors and as Cannataci says[30] referring to a holistic approach has become something of a cliché, but perhaps the context that requires it more than any other is precisely AI.

It is true that there are many different profiles that – given the scope and possible impact of AI – may be affected, but we will try to take a pragmatic and functional approach.

To this end, perhaps the first thing to understand is that there will be roles that will always be necessary, and others that will be contingent depending on the AI system to which they refer. On the other hand, some profiles will be specialised in the subject; and others will be transversal and will contribute their vision from their area of competence. And finally, the intensity of their intervention may also vary: in some cases, they will intervene throughout the entire process; in other cases, they will intervene at specific moments. A good example of a proposal with this grounded vision can be seen in the FRAIA[31] of which, in addition to the various profiles mentioned, the project manager (which is obvious), and the person responsible for the area of knowledge to which the algorithm refers (which we can understand as referring to the area that owns the algorithm if it has a specific purpose) also involves the *legal advisor* in all the phases.

This leads to two reflections: regarding the area that owns the algorithm and its participation, a conflict of interest may arise (also in other roles) due to the obvious interest in the "success" of the system which has (probably)

compliance with the obligations deriving from the carrying out of data protection impact assessments and from prior consultation of the supervisory authority".

[30] Vid Cannataci in his foreword to the work of Mantelero, A. ob. cit. *"For more than forty years, we have been gradually abandoning the mono-disciplinary approach to problem solving to adopt a multidisciplinary approach, often accompanied by an interdisciplinary approach. The perspective gained at the intersection of several disciplines can also be profoundly more accurate and more practical/pragmatic than that constrained by the knowledge and practices of a single discipline. Indeed, the very notion of HRESIA implies taking into account the perspective of other disciplines outside of human rights law, ethics and social impact. Computer science, applied technologies, economics and social psychology are just some of the other disciplines that come to mind and that must be deeply and constantly involved in the way society should think about AI. Talk of 'a holistic approach' has become something of a cliché, but it's hard to think of a context that requires it more than AI... and that's basically the core message of Mantelero's current work".*

[31] The FRAIA mentions different profiles depending on the phases. The profiles it mentions are: Interest Group, Management, Citizen panel, CISO or CIO, Communications specialist, Data scientist, Data controller or data source owner, Data protection officer, HR staff member, Domain Expert, Legal Advisor, Algorith developer, Commissioning client, Project leader, Strategic ethics consultant and Other project team members.

been requested and "interests" the owner area, so that a good practice for managing this conflict, in addition to establishing conflict of interest management rules, is to ensure transparency and document the decision-making process. Another noteworthy aspect in this approach is the involvement of legal counsel throughout the process, which gives us an insight into the importance of legal profiles' participation in human rights-focused impact assessments.

In my opinion, I believe that a data scientist should also be involved in the whole process since, in order to analyse and manage risks that affect rights but involve the use of a technology that is mouldable, it is necessary to know its potential to be "balanced and configured" according to the risk. Of course, the intervention of ethical advisors is a role that, inherited from ethical approaches to AI, has been taking root and some companies have already appointed ethical advisors or committees and approved additional ethical guidelines, usually in line with other international principles. And this will depend on the approach and consequent scope that is agreed for impact assessment, whether it focuses only on human rights or embraces ethical aspects as well.

Another aspect to consider is whether we are dealing with an AI system for internal use (where we should consider internal roles and areas, without prejudice of course to the possibility of relying on external advisors and we should always consider the interested parties, as indicated in Recital 64[32] in line with the GDPR[33] , and which – in terms of AI and in terms of ISO / IEC 42001:2023 and in line with ISO/IEC 22989:2022 – defines them as the *"person or organisation that can affect, be affected or be perceived to be affected by a decision or activity"*.

In addition, and to complete the picture of those involved, it is necessary to consider whether it is an AI system or a product for clients. In these cases, in addition to the obligations set out in the AIA in the case of products, and with regard to the profile of those involved in the EIIDDFFIA, it is required to involve the necessary profiles depending on the product and recipients. A graphic example is the famous case of Hello Barbie[34], although it has little to do with the above-mentioned assumptions regarding the requirement for implementation, but it does serve to illustrate, for example, that if an AI

[32] Recital 64a of the AIA states that, in identifying the most appropriate risk management measures, the provider shall document and explain the decisions taken and, where appropriate, involve external experts and stakeholders.

[33] Article 35.9. GDPR: *"Where appropriate, the controller shall seek the views of data subjects or their representatives on the intended processing, without prejudice to the protection of commercial or public interests or the security of processing operations"*.

[34] Mantelero, Alessandro, in ob. cit p.61 cites the real-life example of Hello Barbie. It

service is to be set up in the field of early childhood education, it is possible that psychologists and/or educational psychologists, for example, will have to be involved.

2.3. When does it take place?

Article 27.1 of the AIA provides that it must be conducted *"prior to deploying a high-risk AI system"* and paragraph two adds that it *"applies to the first use of the high-risk AI system"*. In some models, such as the Canadian RIAT, there is a further check of the assessment prior to deployment,[35] and perhaps, although this seems logical, it could have been underpinned in the legal text, since ultimately the FRIA defines controls that we need to verify before deployment.

Furthermore, Article 27.2 adds*: " If, during the use of the high-risk AI system, the deployer considers that any of the elements listed in paragraph 1 has changed or is no longer up to date, the deployer shall take the necessary steps to update the information"*.

For the purposes of taking into account what constitutes factors, Recital 96 and Article 27.2 clarify that the factors to be taken into account if they have changed are those indicated in paragraph 1 of the same article:

1. The intended uses or purposes.
2. The period of time and frequency of intended use.
3. The categories of natural persons and groups likely to be affected by its use.

was an interactive doll produced by Mattel for the Anglo-Saxon market, equipped with voice recognition systems and AI-based learning functions, which functioned as an IoT device. The doll could interact with users, but not with other IoT devices. The design goal was to provide a two-way conversation between the doll and children playing with it, including capabilities that enable the doll to learn from this interaction, for example by adapting responses to the child's play history and remembering previous conversations to suggest new games and topics. The doll is no longer marketed by Mattel due to various concerns about the safety of the system and the device. To cite just one of the risks associated with the profiling issue we are discussing here: In frequently asked questions about Hello Barbie ("Q: Can Hello Barbie say a child's name? No. Hello Barbie does not ask for a child's name and is not programmed to respond with a child's name, so it will not be able to recite a child's name"). But Mantelero quotes a response in the dialogue with the doll: "Barbie: Sometimes I get a little nervous when I tell people my middle name. But I'm so glad I told you! What is your middle name? This example shows that, in this case, a team is needed to work on sentences and dialogues, and for this purpose it may be necessary to call in psychologists specialised in early childhood education or similar profiles. To illustrate this example, he cites the following:

[35] The Canadian RIA states: "The RIA should be completed early in the design phase of a project. The results of the RIA will guide the mitigation and consultation requirements to be met during implementation of the automated decision system under the directive.

The RIA should be completed a second time, prior to production of the system, to validate that the results accurately reflect the system that was built".

4. Specific risks of harm that may affect categories of persons or groups of persons.

5. Human oversight measures.

6. The measures to be taken in the case of the materialisation of those risks, including the arrangements for internal governance and complaint mechanisms. In the same vein, ISO 42001:2023, when referring to RRAAs and also to System Impact Assessments, states that they should be carried out *"at planned intervals or when significant changes are proposed or occur"*.

EIDDFFFIAs must in turn be part of an AI Management System, which in the end may or may not be based on a standard such as ISO 420012: 2023, and may in turn be integrated into another management system, but which like all of them is based on a PDCA, Deming cycle or continuous improvement process, as also indicated by the AIA[36] , which – by definition – means continuous improvement and – therefore – updating.

2.4. What is it about? Substantive scope. A preDPIA

When we are going to carry out an RRAA as well as an impact assessment, we must first be clear about what is being done. For example, when we talk about data protection we are all clear that we are talking about processing (without prejudice to the debates on the greater or lesser granularity of this concept); when we talk, for example, about an RRAA within the framework of the NSS for information systems. And following the case of data protection, not all processing operations should have to have an DPIA, so it is necessary to carry out what is colloquially known as a PreDPIA or PrePIA, which analyses the need to carry out the DPIA or not.

The same should be done when we talk about AI systems, we talk about Fundamental Rights Impact Assessments of high-risk AI systems in the entities to which they apply, without prejudice to the fact that other non-obliged entities may voluntarily decide to carry it out.

In this case, there are two parts to the object of the evaluation when assessing whether or not it should be carried out:

(a) High-risk AI systems are involved.

b) That these AI systems refer to the specific entities and/or services within the subjective scope referred to above.

c) And – finally – obviously, that they affect or may affect fundamental

[36] *Recital 65 AIA provides: "The risk-management system should consist of a continuous, iterative process that is planned and run throughout the entire lifecycle of a high-risk AI system. That process should be aimed at identifying and mitigating the relevant risks of AI systems on health, safety and fundamental rights. The risk-management system should be regularly reviewed and updated to ensure its continuing effectiveness, as well as justification and documentation of any significant decisions and actions taken subject to this Regulation".*

rights. What an AI system is, which ones are high-risk and who are the obliged entities has been dealt with in other chapters of this book, but it deserves more attention to focus on what it means that they may affect fundamental rights. On the one hand, it is important to point out that, without prejudice to other approaches that can be taken "further" by adding ethical or social aspects, the minimum basis for evaluation must be fundamental rights; and, therefore, the source of requirements must come from the Charter; if you like, harmonised with the fundamental rights contemplated in the Spanish Constitution, of almost absolute coincidence as we have mentioned. It is also important to clarify that, in the European sphere, unlike the UDHR, and thanks to the specific recognition by the ECHR, the inclusion of environmental protection is clear, although in the case of Spain it would not be included as a fundamental right, since the mention of environmental protection is made in Chapter III of the Constitution, and there is a debate about it[37], perhaps because of this lack of homogeneity at the European level, the AIA "settles the debate" about its inclusion. But there is also a close relationship between the possible impact of AI on the environment and vice versa[38], which is confirmed by the various references to the environment in the AIA and in some of the ISOs that deal with AI.

This need to delimit the scope of the possible rights affected by the AI system will require us to carry out a PreAIA (pre AI Assessment), in a similar way to how we have been carrying out PreDPIAs in the field of data protection.

At this stage, perhaps we can not only identify the rights that are affected, but also evaluate if the significance or significant influence of any of these

[37] A possible discrepancy of the scope of rights in the Spanish Constitution and European context could be argued, delving into whether, in the Spanish case, we should extend such analyses to the whole of Chapter II of Title I or only to section 1ª , and/or if so, to strike a blow against the EU Charter of Fundamental Rights + ECHR. If one assumes that the environment is included, then, in the part of Article 9 "risk management" where it talks about environment/health and then DDFF, perhaps one should focus on integrating more "regulated" environmental RRAA either in the context of specific sectoral/administrative/ISO etc. requirements.

[38] As the UN Special Rapporteur on human rights and the environment says "All human beings depend on the environment in which they live. A safe, clean, healthy and sustainable environment is indispensable for the full enjoyment of a wide range of human rights, including the right to life, health, food, water and sanitation.

In the absence of a healthy environment, we are unable to realise our aspirations. And we may not even be able to meet the minimum standards of human dignity." https://www.ohchr.org/es/special-procedures/sr-environment/about-human-rights-and-environment Last accessed on 12/03/2023.

rights in the project warrants a specific impact assessment, separate from the overall FRIA. However, in my opinion, we should strive to conduct integrated FRIAs, taking into account the potential exception -perhaps more common- of data protection and information security rights[39], which are already well developed and have mature roles, and which, although we will comment later, should be carried out "jointly" and in coordination, but this does not necessarily mean that they should be "fully" united.

Another aspect to consider, in order to make the analysis of these rights feasible, is grouped by areas which, as we will see later, can also be linked to the groups or groups impacted[40].

There is no doubt that when conducting an AI in which so many rights may be affected, an analysis of various sources of requirements and their consequent controls for each of the rights must be carried out, which places those conducting the AI in a titanic task, since talking about rights implies knowing not only their description but also their grounding in implementing legislation, guidelines and resolutions of administrative and judicial authorities etc., even with the help of a multidisciplinary team as we have mentioned.

Although the AIA has maintained a vision centred on fundamental rights without including ethical and social aspects as mandatory, an aspect that has been criticised by part of the doctrine[41], the fact is that these can and will be added, especially in organisations that – beyond human rights – are already addressing such issues.

In addition to all of the above, it is necessary to add, as is usual in large corporations, the internal regulations on these matters, which cannot be subtracted, but which can increase requirements and their corresponding controls.

[39] In fact, ISO 42001:2023 sets them as an example.

[40] For example, Telefónica's HRIA, which can be consulted at https://www.telefonica.com/es/sala-comunicacion/reportes/el-proceso-de-debida-diligencia-de-telefonica-en-ddhh-y-medioambiente/amp/, groups them into the following 5 areas: Ethics and Governance, value chain, operations, human resources and products and services.)

[41] Mantelero, A. in ob. cit p. 173. already said *"... after several years of debate on the ethical dimension of AI, the prevailing view seems to be to delegate ethical issues to other initiatives not integrated in the legal assessment. Just as focusing exclusively on ethics was critical, this lack of integration between the legal and social impacts of AI is problematic. An integrated assessment model, such as HRESIA, could overcome this limitation in line with the proposed risk-based model".*

V. Steps to be taken in a FRIA

A FRIA is itself a process that has several phases laid out in a PDCA, which in turn can be integrated into the action plan in an Artificial Intelligence management system; and this, in turn, can and usually will be integrated into other management systems, as we have already mentioned. However, here we will look at the PDCA itself, which constitutes the DPIAFFIA itself. Basically, the various Impact Assessment Methodologies in which there has been a "tradition" (for example, those of Human Rights or the DPIA) have a similar, widely accepted structure, without prejudice to any nuances that may exist: either because of who proposes it or because of the subject matter being assessed.

It is also important to note that FRIA, as with HRIAs, could be carried out in an integrated manner or not[42] with other impact assessments, which, as with the former, also has advantages and disadvantages.

It should also be considered that there may be different examples of AI systems: from examples dedicated to specific purposes (AI scenarios that support specific processes or areas) to others related to more complex systems (such as *Smart Cities*). This may mean that, although the same methodology is used, and depending on the size of the AI system, adjustments or "additions" may have to be made. For example, in cases where many AI systems are combined, an additional Impact Assessment may have to be carried out in order to have an overall picture.

The content required by the AIA will be included in each of the proposed phases, given that, in order to be available, it must be extracted from them.

As we have already mentioned, and given that the FRIA itself is a PDCA, the first phase is the Plan phase. The Plan phase, in turn, is made up of different phases that will lead to the FRIA Report where, among other issues and as we will see, the corresponding risk treatment actions will be defined.

Phase 1: Preliminary analysis of the need for a FRIA and specification of the systems and rights affected (initial scoping).

First, we will have an inventory of the systems within the scope. As for the systems that should be subject to FRIA, they are the high-risk ones, and this issue has already been addressed in this book.

As regards the fundamental rights on which the assessment should focus, we have also indicated in this chapter which ones are concerned. This analysis

[42] See section A.6.8. entitled "Should HRIAs be independent or integrated?", p. 27 of https://www.humanrights.dk/sites/humanrights.dk/files/media/document/HRIA%20Toolbox_INTRO_Spanish.PDF last consulted on 13/03/2024.

can be carried out on the basis of knowledge of the AI system, the AIA and at the operational level with a *checklist* without much fieldwork and stakeholder involvement.

Phase 2: Context, planning and detail within the scope

Before starting a FRIA, we must have the necessary information. In different methodologies, this is done in different ways[43].

In our opinion, the main aspects to know are the following:

1. Determine the team that is going to carry it out, which has to do with determining which of the possible interveners referred to in this chapter are required in this case, as some profiles will always be necessary and others will depend on the case.

2. Determine the detail within the scope of the FRIA, going deeper into certain aspects from the previous scope established) which implies:

a) Knowing the type of project or activities of the organisation that are affected by the AI system.

b) Have a sufficient understanding of the AI system, particularly of the information it holds in relation to the rights potentially affected, given that the following aspects should be reflected in the relevant report:

(i) a description of the implementer's processes in which the high-risk AI system will be used in accordance with its intended purpose;

(ii) a description of the time period and frequency with which each high-risk AI system is expected to be used

(iii) the categories of natural persons and groups likely to be affected by its use in the specific context;

[43] For example, in the case of the Canadian RIA it is said that before starting it is useful to have information about: ITSM business practices the management decision that will inform or be made by the automated decision system; the context in which the system will be used and how the system will assist or replace the judgement of a human decision maker the clients subject to the decision, including evidence of any vulnerabilities (e.g. socio-economic, demographic, geographic); the algorithm, including the parameters and data processing techniques, and the output; the input data used by the system, including the parameters and data processing techniques, and the output; the input data used by the system, including the data processing parameters and techniques, and the output; the input data used by the system, including the data processing parameters and techniques, and the output;the algorithm, including data processing parameters and techniques, and the output; the input data used by the system, including details on type, source, collection method and security classification. The system has either planned or implemented quality assurance measures, and it plans to communicate information about the initiative to both customers and the public through transparency measures. The system should consult both internal and external stakeholders and keep a record of any recommendations or decisions it makes, along with any documentation or justifications it generates for these records..

c) Have a correct knowledge of the context.

d) Identify relevant stakeholders: stakeholders in the narrow sense, as well as rights-holders, duty-bearers, other relevant parties.

3. Determine the methodology to be used and the sources of requirements to be used.

Also, as mentioned above, this obligation to perform the FRIA is on the implementer, but the implementer may rely on those carried out by the deployer (ex. article 27 2. AIA, as mentioned above), so that this will also be part of the knowledge of the AI system.

Finally, although not required by the AIA, and given that we are considering the possibility of conducting an DPIA in conjunction with a FRIA, at least a description of the personal data processed should be available[44].

Phase 3: Necessity, proportionality and quality of data

As is well known, data protection impact assessments incorporate an analysis of the necessity and proportionality of the processing. This aspect has recently become controversial, given that the AEPD in its new Guide on presence control processing using biometric data[45] changes the criteria followed to date and by "tightening the sense" that personal data should only be processed if the purpose of the processing could not reasonably be achieved by other means, as indicated by colleagues Patricio Monreal and Maria Loza *"it will be practically impossible to overcome the judgement of necessity of the processing (except in very specific and residual cases), as there will always be other less intrusive and equally effective means"*. And they add, in relation to such processing involving AI, that the aforementioned Guide itself indicates that *"they must take into account the prohibitions, limitations and requirements established in the regulations on Artificial Intelligence", but will it ever overcome the judgement of necessity? If we turn to the AEPD document*[4] *"*Adaptation to the GDPR of processing operations that incorporate Artificial Intelligence. An introduction*", it states that the use of AI-based solutions may entail a high level of risk and therefore "it should be assessed whether the purpose of the processing cannot be achieved using another type of solution that achieves the same functionality, with an acceptable performance margin and a lower level of risk".*

Therefore: obviously, high-risk AI systems processing personal data should at least entail this analysis of necessity and proportionality in relation to the processing of personal data[46], but the AIA does not make a reference to this aspect. However, in my opinion, as has been done by some method-

[44] What the DPIA of the AEPD in its guidance (cite) calls "Describing the data lifecycle".

[45] https://www.aepd.es/documento/guia-control-presencia-biometrico.pdf

[46] As required by Article 35.7.b of the GDPR.

ologies[47], it does require that there be a "moment" in which necessity and proportionality are analysed, assessing the aspects that have been taken into account to implement the AI system, considering aspects such as: why precisely that AI and not another; the consequences of not implementing it; and, at least preliminary (without going into the risk analysis that will be the subject of in-depth work), that at least a prior approximation has been made of the benefits and sacrifices that it entails. Obviously, this balance or weighing differs greatly between the public and private spheres, which is why this analysis in the case of FRIA, which is always required in association with public services, makes even more sense. An example in the public sphere is given by the aforementioned FRAIA: *"Suppose, for example, that an algorithm is an eminently suitable and necessary tool to improve the efficiency of decision-making, but there is a real risk that the tool reinforces discriminatory patterns. In that case, is it reasonable to continue to implement the tool? It is not possible to formulate strict and objective criteria to determine the weight and balance of various rights, interests, objectives and public values. Generally we can say, however, that the more serious the expected infringement of fundamental rights, the more serious the social objectives will weigh in comparison.*

[47] The AIA gives some examples of possible reasons for introducing automation:

1. Existing backlog of work or cases.
2. Improve the overall quality of decisions.
3. Lower transaction costs of an existing programme.
4. The system is performing tasks that humans could not perform in a reasonable period of time.
5. Use innovative approaches.
6. others that may be specifiedAlso the FRAIA in its Part 1 deals with the "Why" of the intention to develop, purchase, adjust and/or use an algorithm (hereafter abbreviated: the use of an algorithm). What are the reasons, the underlying motives and the intended effects of the use of the algorithm? What are the underlying values that drive the deployment of the algorithm? These general questions should first be discussed in a decision-making process about the use of algorithms, before eliciting questions about, for example, preconditions or possible impact on fundamental rights issues. The answers given to the questions in this part are relevant for answering the more specific questions in the following parts.

And then in Part 4 when analysing fundamental rights it says: *Necessity and subsidiarity.* A wide range of tools and means can be used to achieve policy objectives, including algorithms. Even if a specific tool is chosen, it can often be used in a variety of ways. In addition, it may sometimes be possible to soften the harmful effects of a given instrument through compensatory or mitigating measures. The choice that can be made between various tools is central to the question of the necessity and subsidiarity of choosing a specific algorithm.

Balance of interests/proportionality.

Even if an algorithm seems to be an adequate and necessary tool to achieve the formulated objectives, a final step is always necessary. This step has to do with the relative weight of the fundamental right at stake, compared to the relative weight of social objectives and public values.

As a result of my reflections with Jordi Morera, I raise two possible debates that intersect with privacy:

1. If an analysis of a High Risk system is being carried out and the authority has been notified following the FRIA: should it be understood that this weighing is automatically exceeded in the DPIA of the processing that is the object of said AI system? Perhaps the answer should be in the affirmative, since it would not make sense for an administrative authority (the AEPD) to make a pronouncement against the deployment of an AI system, which, although it is High Risk, has been recognised as such in the AIA and which has also been notified (and validated) by the authority, at the risk of conflict between different authorities. To accept the contrary would be to attribute to the AEPD the function of a "negative legislator", since if it indicates that a processing operation is not proportionate due to an AI system that the AIA has not prohibited and which it regulates with a series of guarantees (high-risk ones), it would be exceeding its powers and correcting the legislator itself.

2. Notwithstanding the above, there is another point that should give the DPOs reflect: if a High-Risk AI system is used, where personal data is processed, we should automatically assume that a DPIA is required, because after all, the AIA legislator has determined that such a system generates a high risk for the fundamental rights, and – therefore – its assessment can be extrapolated to the personal data used for it. In other words, when it comes to a high-risk AI system, it can be understood as automatically equivalent to high-risk processing and consequently to the mandatory conduct of a DPIA.

Phase 4. Risk management

We have already referred to the risk approach of AIA in several sections of this chapter and it is the subject of another chapter in this book, so we will not dwell on it here. Risk management is at the heart of any AI, and therefore also of FRIA. For this reason, and even taking into account the differences we have discussed on the scope and meaning of article 9 AIA with respect to risk management and the risk part mentioned in article 27, risk management has a broadly consolidated backbone. For this, in addition to taking into account the specificities of the AIA, and taking into account the AIA, it is best to draw inspiration from criteria accepted as global standards. In this case ISO 31000:2009 Risk Management – Principles and Guidelines, from which the rest of the ISO standards are inspired and adapted to specific environments, in addition to the specific ISO in AI to which we have also referred, particularly ISO/IEC 23894:2023*, guidance on risk management in AI*.

The phases to follow, with nuances (AIA groups identification and analysis), coincide, and we can say that they are:

A) Identification of risks

The aim is to identify the known and reasonably foreseeable risks the AI system may have on the fundamental rights that have been defined in the scope, when it is used in accordance with its intended purpose.

B) Analysis of possible risk scenarios

The vulnerability of the AI system to these rights must then be estimated in two senses: probability and impact.

Although Article 7 of the AIA does not refer to RRAAs and FRIAs, but mentions criteria to be taken into account by the Commission in assessing the modification of systems considered as high risk in Annex III, it may be possible to also consider these elements for the assessment of likelihood and impact.

a) According to ISO, probability refers to the possibility of an event occurring. It can be defined or determined objectively or subjectively, qualitatively or quantitatively, and described using general or mathematical terms (such as a mathematical probability or a frequency in a given period of time). This is one possible table for calculating the likelihood of an AI system affecting each risk factor identified for each human right in the scope, although there may be other valid ones:

NAME	DESCRIPTION
(VERY LOW) >= 1 time every 100 years	Occurs at least once every 100 years
(LOW) >= 1 time every 10 years	Occurs at least once every 10 years
(AVERAGE) >= 1 time per year	Occurs at least once a year
(HIGH) >= 10 times a year	Occurs at least 10 times a year
(VERY HIGH) >= 100 times per year	Occurs at least 100 times a year

b) On the other hand, the impact or consequences of the defined risk scenarios materialising can be certain or uncertain and can have direct or indirect effects on the objectives. Likewise, and according to the aforementioned ISO, it can have positive or negative effects.

Consequences can also be expressed qualitatively or quantitatively. As in the case of probability, various scales can be used, but one possible way of representing the possible severity of the consequences that, for the analyzed fundamental rights, would arise from an event affecting the AI system would be the following:

IMPACT	**DESCRIPTION**
DESPRECIABLE	Right holders will be largely unaffected or will encounter some minor inconvenience.
LIMITED	Right holders may encounter non-significant inconvenience
SIGNIFICANT	Right holders will encounter significant consequences that they should be able to overcome without serious difficulty.
MAXIMUM	Right holders will encounter significant or even irreversible consequences, which cannot be overcome.

C) Estimation and assessment of risk scenarios or hazards

As the AIA (Art. 9.2.b.) says for RRAAs in general and is applicable to this phase of FRIA, the next step is to estimate and assess the risks *"that may emerge when the high-risk AI system is used in accordance with its intended purpose, and under conditions of reasonably foreseeable misuse"*. The reference in point (c) to "*evaluation of other risks possibly arising, based on the analysis of data gathered from the post-market monitoring*

system referred to in Article 72" and obliging providers to establish and document a post-market surveillance system in proportion to the nature of Artificial Intelligence technologies and the risks of high-risk AI systems, does not impact the initial assessment but rather underscores that the methodology requires iterations.

The formula used for risk estimation is RISK = PROBABILITY x IMPACT (consequence)**.** In this way, a risk matrix is generated, which, coherently with the defined thresholds (we insist that other valid thresholds may be used), could be as follows:

	Impact			
Probability	**Despicable**	**Limited**	**Significant**	**Maximum**
Very high	Medium	High	Very high	Very high
High	Medium	Medium	High	Very high
Average	Under	Medium	Medium	High
Low	Under	Under	Medium	Medium
Very low	Very low	Under	Under	Medium

The result of the assessment is the initial risk.

D) Risk management or treatment

Given the initial risks identified, one of the following strategies or ways of dealing with the risk, which are also specified in Article 9.4 of the AIA, should be adopted, in line with the above-mentioned risk management theory:

1. One option is to mitigate or reduce the initial risk by implementing controls that reduce the risk below the threshold defined as acceptable. Two options can be adopted for this purpose:

a) Reduce the impact caused by a risk scenario.

b) Reduce the likelihood of a risk scenario materialising.

2. Another option is to avoid or eliminate the risk by nullifying, excluding or replacing the design element or functionality. This option is not always feasible as it can sometimes result in the loss of essential functionality.

3. The last theoretical option, according to "general risk management theory" is to ignore or assume the risk. That is, to do nothing to deal with it. Theoretically this would be possible in three scenarios:

1. Where the impact or consequence is acceptable.

2. Where the risk is acceptable.

3. And when the cost of the measures to be adopted is disproportionate to the impact and risk. But: is it possible to take acceptable risks on fundamental rights?

We have to start from the fact that the AIA has already done prior regulatory risk management work by prohibiting certain AI systems. Moreover – as is well known – there is no such thing as "zero risk".

According to Mantelero, unlike the notion of acceptable risk which *"comes from product safety regulation*[48] *in the field of fundamental rights the main risk factor is proportionality and implies the absence of risk or minimal risks"* and concludes that *"if we accept this interpretation, acceptability is incompatible with the high risk of adverse impacts of AI on fundamental rights and any impact assessment based on a quantification of risk levels will play a crucial role in risk management"*. It is true that if the risk equation is composed of two factors (likelihood and impact) and the impact on fundamental rights we consider to be always high, maintaining such a position would lead us to flatly deny the use of the risk management methodology advocated by the AIA. However, Article 9.4 of the AIA itself overrides this

[48] Mantelero, A., in ob. cit. p 172 *"Article 2(b) of Directive 2001/95/EC on general product safety defines a safe product as one which presents no risk or only the minimum risks compatible with the use of the product, considered acceptable". Indeed in this regard Recital 27 of the AIA states that "High-risk AI systems should only be placed on the Union market, put into service or used if they comply with certain mandatory requirements. These requirements should ensure that high-risk AI systems available in the Union or*

view and thus validates the existence of acceptable risks by stating that risk management measures *"shall be such that the relevant residual risk associated with each hazard, as well as the overall residual risk of high-risk AI systems, are considered acceptable"*. It is also worth remembering that we are already carrying out legally required RRAAs and DPIAs based on a fundamental right such as the protection of personal data and using the risk management methodology. The AEPD in its guidance has said that *"low and medium levels of residual risk that will require proportionate management efforts throughout the lifecycle of the processing could be considered as acceptable residual risk levels"* and cites the WP Guidelines248 with examples of unacceptable risks[49], but obviously there are many other acceptable risks. In other cases, such as the right to the environment, this can be seen more clearly with an example, such as a construction that causes a lot of noise and where an acceptable noise threshold is set that takes into account legislation (there is often local regulation) or other criteria such as environmental impact studies where necessary, or even community consultation. But measures such as less noisy machines, noise barriers, limiting the hours of noise etc., could be taken to lower the initial risk to the acceptable noise threshold. But let's take an AI example: Let's imagine an AI system applicable to recruitment processes where we analyse the risk of breaching the right to equality. The system in the learning phase is trained with CV data to select the most suitable CVs to follow the process. A possible risk identified for equality is the bias of the algorithm (let us imagine that it takes more data from men who pass the process in the training) that could lead to discriminating against women in the inference phase. One possible way to determine the acceptable threshold is to consider a percentage of false positives and negatives as the maximum acceptable. If so, measures could range from re-evaluating the training data or modifying the algorithm to balance or compensate for this "risk", before going to the extreme measure of not using such an AI system.

In short, assuming as *lege data* (current law) the use of the risk management methodology established by the AIA and assuming also that high impacts on fundamental rights cannot be tolerated, there would be no acceptable high-level risks.

the results of which are otherwise used in the Union do not pose unacceptable risks to important public interests of the Union recognised and protected by Union law'.

[49] "An example of unacceptable high residual risk includes cases where data subjects may face significant or even irreversible consequences from which they cannot recover (e.g., illegitimate access to data that poses a threat to the life of the data subjects, redundancy, financial danger) or where it appears obvious that a risk will exist (e.g., failure to reduce the number of people accessing the data because of its modes of sharing, use or distribution, or where a known vulnerability is not corrected).

Well, depending on the nature of the risk, safeguards or controls must be adopted, which may incorporate measures to lower the initial risk to the threshold of acceptable risk, measures that may be of different types, but without doubt and unlike the controls required of the provider according to the risk analysis of Article 9, where there is a greater weight of technical controls, in the case of the controls derived from the FRIA (the controls relating to Governance and legal compliance that correspond to the implementer will have more weight).

Article 27 1. refers to a number of specific measures to be taken into account: Human oversight measures, arrangements for internal governance and complaint mechanisms.

In any case, recital 64 states that the measures to be taken *"should take into account the generally acknowledged state of the art on AI, be proportionate and effective"*.

As we have indicated and as a reference of possible measures, in addition to those indicated in the AIA, we can consider those referred to in the ISOs mentioned above.

As a result of treating the risks, the residual risk is obtained, defined as the level of risk resulting from the treatment once control measures have been applied to mitigate and/or reduce its level of exposure in relation to the set of risk factors identified. Unlike inherent risk, residual risk takes into account the control measures defined on the AI system. Therefore, as stated in Article 9.5. AIA: *"Risk management measures... shall be such that the relevant residual risk associated with each hazard, as well as the overall residual risk of the high-risk AI systems is judged to be acceptable"*.

Ultimately, the conclusion to be reached by the FRIA is whether, given the initial risks, by applying the appropriate measures or controls, we will be able to bring the residual risk below the acceptable risk.

These measures should be implemented in the OD phase, and reviewed and continuously improved in the *check and act* phases.

But before the OD phase, the AIA adds that they must not only be implemented, but also tested, which is discussed in other chapters of this book.

Another important aspect of FRIA as part of the risk management system is that, as Article 9 1. states, it *"shall be established, implemented, documented and maintained*". In other words: documentation and maintenance is essential.

F) Visual representation and management

As we have indicated, we are dealing with an AI that may consider in its scope, depending on the AI system, several affected rights, which, multiplied by the possible risk factors identified and assessed, may entail numerous risks to be addressed and consequently many controls to be applied. This leads to several critical issues to be addressed:

1. One of them is the clear representation of the data. It is not just a matter of producing a report that contains the content required by Article 27, but that the report must be understandable, and when we are talking about so much data this is not always the case. The truth is that many existing HRIAs have had a lot of literature, but the results are not graphically visible, others are. For this, graphs can be very useful. One proposal could be the radial graph[50], but there are other models, for example a typical heat map could be used for each right and aspect or dimension analysed (e.g., security or others), in which the rights are identified and another as a summary in which the risks for each of them are visualised; and additionally a consolidated one (especially suitable when we are talking about many affected rights) in which a global vision is available[51].

2. Another aspect has to do with the need to use tools (GRC tools -Governance, Risk and Compliance- have become fashionable) that allow not only the RRAA to be carried out but also to integrate it into the PDCA, allowing support to be given to the dynamic vision consistent with the improvement plan that it requires, integrating it into measures required by other regulatory frameworks and other management systems.

G) Communication to the authority

As stated in Article 27.3 and after the FRIA has been carried out*, "the deployer shall notify the market surveillance authority of its results, submitting the filled-out template referred to in paragraph 5 of this Article as part of the notification"*.

The existence of a template that the authority makes available to those responsible for reporting obviously conditions the information and the format of the report, in order to comply with this obligation; but – in our opinion – we cannot confuse the obligation to report to the authority through a questionnaire with the evaluation itself, the report itself and the need for management that it will require.

In short, the EIDDDFFIA should contemplate the information and format of the questionnaire to be communicated to the authority, but foreseeably with more information and in a system that allows iteration and continuous improvement.

This obligation applies to all those responsible for high-risk AI systems except those covered by Article 47.1 of the AIA which are also exempted from the conformity assessment procedure and concerns market surveillance

[50] According to Mantelero, A. in ob. cit. p.59, *"the radial graph is therefore the best tool to represent the outcome of the HRIA, showing graphically the changes after the introduction of mitigation measures"*.

[51] Such a vision would be in line with the possibility referred to by the AIA in several articles of having a vision of global impact.

authorities placing AI systems on the market or making them available in the EU for exceptional reasons of public safety or protection of human life and health, protection of the environment and protection of key industrial and infrastructural assets.

H) Communication to stakeholders? Publication?

As mentioned above, the AIA includes the need to notify the market surveillance authority of the results of the assessment by means of a template. But the question is, should it be published? The AIA does not require this. It is obviously necessary to ensure that the decisions of the AI systems are fair and impartial. We can consider that this is guaranteed by the fact that it is communicated to the authority; however, if the aim is also to achieve trust and in the case of the public sector also citizen participation, perhaps publication should be considered as good practice (given the lack of legal requirement), which could contribute to the improvement of systems in aspects such as the reduction of biases. Indeed, it is "curious" that the AIA mentions the involvement of stakeholders as well as experts in identifying the most appropriate risk management measures, but does not oblige them to be informed of the outcome.

As it is possible that some information may not be published for reasons of business confidentiality, a summarised or redacted version could be published.

VI. Fundamental rights impact assessments and data protection impact assessments

It is not the purpose of this section to analyse the multiple intersections that can occur between AI systems and the processing of personal data, which are dealt with in another chapter of this work. It is not even the purpose of this section to develop the methodology of the specific PDIAs on which, as there is abundant literature and a "well-established tradition in recent years", and to which reference has been made in part in this chapter. The purpose of this section is to refer to the provision in Article 27.4 AIA specifically that, if any of the obligations set out in that Article is already fulfilled by the data protection impact assessment carried out under the GDPR, the fundamental rights impact assessment shall be carried out together with the data protection impact assessment.

As we have already advanced throughout this chapter and as is well known, data protection and also DPIAs are one of the types of impact assessments that have been most widely deployed in recent years in Europe. The AEPD has not only drawn up several guides on the subject and has issued reports and various sanctioning resolutions – some of them controversial – on the matter.

To speak of joint assessments between the DPIAs and FRIAs is first to determine that: On the one hand, we are dealing with an AI system; and that, likewise, it processes or will allow the processing of personal data. As stated in the document published by the AEPD *"Compliance with the GDPR for processes that incorporate Artificial Intelligence. An introduction"*[52], *"If an AI component carries out the processing of personal data, draws up profiles on a natural person or makes decisions about a natural person, it will have to be subject to the GDPR. Otherwise, it will not be required"*.

There are numerous examples of AI systems that process personal data, including facial recognition, biometric data processing, recruitment, and marketing, among others. However, there are also other types, such as those related to industrial quality processes, that do not process personal data.

Furthermore, it is not always easy to determine whether or not personal data will be processed at any stage of the lifecycle of an AI system.

As we have advanced, in the proposed methodology, there is a prior phase or Pre-impact assessment to delimit the scope of the rights affected. In this phase it will be determined (*prima facie,* and without prejudice to the fact that if it is subsequently detected that there is data processing or not, the position adopted will be "rectified") whether there is data processing and also whether there is an obligation to carry out a DPIA, for which – as is well known – there are already not only criteria established by the GDPR that have been interpreted and developed[53]. A question that could perhaps be raised here is whether we can "automatically" assume that if there is a processing of personal data (whether or not it is one of the cases where there is an obligation *per se* to carry out a DPIA) and it is also a High Risk AI, whether this is equivalent to a necessary PDIA.

It is essential that the DPO is involved in the analysis and decision making process, when it has been appointed. In the event that the DPO has not been appointed because it is not mandatory or has not been appointed voluntarily, the person who provides legal advice on data protection and the CISO should be heard on the matter, as suggested in the aforementioned AEPD guide on DPIA, adding that such *"suggestions must be recorded in documents, as well as the decisions taken on the basis of them"*. Even in cases where there is no obligation to do so, the data controller may decide to do so for various reasons cited in the

[52] https://www.aepd.es/documento/adecuacion-rgpd-ia.pdf

[53] See footnote 9 of this chapter.

aforementioned guide, such as: *"in the interests of greater diligence in implementing proactive responsibility"*, or to *"improve the quality of its products and services, promote the culture of data protection in its organisation or simply as a mechanism to ensure the trust of its customers"*.

But, to continue with the analysis, supposing that we are faced with an AI system obliged to carry out a FRIA and that there is a processing of personal data that is obliged to carry out an DPIA, or in both cases that it is voluntarily decided to carry them out, what the aforementioned article 27.4 AIA states is that the FRIA will be carried out jointly with the DPIA. It is logical that they should be carried out jointly, since privacy, like security from the design stage, and in general compliance from the conception of any process or system, is a process that should tend to be integral and integrated. However, data protection has been a subject that -as we have already indicated- has important specificities, some of which -such as the need and proportionality to which we have referred- have a different meaning and depth (according to the interpretation given by the data protection authorities), and likewise the catalogue of risk scenarios and corresponding controls is very detailed, which are already integrated in a data protection management system, often already integrated in an ISMS, as well as other specificities. Therefore, in my opinion, one thing is that they should ideally be carried out jointly, and another is that it is possible and sometimes advisable that, without prejudice to the coordination of both assessments, the data protection assessment should become an *ad hoc* report itself (with the corresponding "calls" or references from the "FRIA" report). Indeed, the AIA itself in Annex B of Section VIII requires that there should in any case be separate summaries in the case of the registration of high-risk systems:

A) On the one hand, a summary of the conclusions of the fundamental rights impact assessment carried out in accordance with Article 27;

(B) and on the other hand a summary of the data protection impact assessment carried out in accordance with Article 35 of Regulation (EU) 2016/679 or Article 27 of Directive (EU) 2016/680, as specified in Article 27.6 of this Regulation, where applicable.

Another aspect to consider in the execution of these joint scenarios is that in conducting the DPIA, the provider should always be asked what measures it has implemented and rely on the information in the RRAAAI.

Obviously, when we are talking about mature and integrated systems, the report will be one thing and the management of the identified risks will be another thing.

VII. Recapitulation and conclusions

We have framed impact assessments as tools for weighing up fundamental rights "imported" from the Anglo-Saxon tradition and which are carried out by the subjects themselves who are part of the decision-making process, without prejudice to those carried out by the legislator, the courts and the administration in certain cases.

We have then taken the example of data protection to distinguish conceptually between Data Protection Risk Analysis (RRAA) and Impact Assessments (DPIA), although we have concluded that these differences cannot be extrapolated exactly to this area, since – as also discussed in the chapter on Pere Simón – the risk management of Article 9 of the AIA refers to all high-risk AI systems and is an obligation that providers must comply with; In contrast, the impact assessment provided for in Article 27 (FRIA) is oriented towards specific cases and the parties obliged to carry it out are those responsible for the deployment of these specific high-risk systems.

Various typologies of impact assessments such as HRESIA or HRIA (Human Rights, Ethical and Social Impact Assessment), PIA (*Privacy Impact Assessment*) or DPIA *(Data Privacy Impact Assessment*), SIA (*Social Impact Assessment*) and *EtIA (Ethical Impact Assessment*) have been analysed, as well as the background of fundamental rights impact assessments of AI systems prior to AIA, with references to methodologies used in different countries and proposed by different organisations.

In the following section, we have looked at the FRIA in the text of the AIA and analysed its evolution, processing, and final content with respect to the new article 27, which has been a "recent introduction" since it appeared in the parliament's version, in order to subsequently focus on how it has been regulated and to carry out a methodological approach, of which we can highlight the following aspects:

1. Those obliged to carry out FRIAs are only certain specific deployers.

2. It is important that the team involved is multidisciplinary and that it is necessary to consider that there will be roles that will always be necessary and others that will be contingent; as well as profiles specialised in AI and others that will contribute their vision from their field of competence. It is also important to involve stakeholders and potential experts in its implementation.

3. The FRIA should be conducted both prior to deployment of the particular high-risk AI system and should be updated in the event of changes in factors relevant to the AI system (uses or purposes, time period and frequency of intended use or categories of individuals or groups affected), risks that were identified or measures that were taken to manage those risks.

4. The substantive scope needs to be narrowed down, prior to implementation, by means of a preFRIA, taking into account the following requirements:

a) High-risk systems are involved.

b) Such systems relate to the specific entities and/or Services within the subjective scope indicated in the AIA

c) And – finally – that it affects fundamental rights, focusing on which ones.

5. Taking into account this previous phase of narrowing the scope, we could establish a proposal for the phases of the FRIA itself, which would be as follows:

Phase 1. Preliminary analysis of the need to carry out the FRIA and specify the Systems and rights affected.

Phase 2. Knowing the context (which goes beyond the preFRIA phase, as it will involve going deeper into different aspects such as the implementer's processes, the expected period and frequency of use or the categories of individuals and groups affected in the specific context) and planning it, which will also entail landing the corresponding timeline with milestones and the team involved in each phase in relation to a methodology to be used.

3. Step 3. Although not specifically provided for in the AIA, in our view an analysis of the necessity and proportionality of the deployment of the AI system should be carried out, not only when required in relation to the processing of personal data under the GDPR.

4. Phase 4. The risk management phase, which includes the identification of risks, analysis of possible risk scenarios, estimation and evaluation of risks and subsequent management by proposing appropriate measures. At this point, we have paid special attention to the assumption of acceptable and assumed risk lege data that the AIA assumes that there is no such thing as 0 risk, that there may be tolerable risks, but – this is a personal opinion – we understand that given that we are talking about one of the risk factors being impact, and that we are talking about impact on fundamental rights, the threshold of acceptable risk should be lower and high risks should not be accepted.

6. That all information gathered and analysis performed will lead to a conclusion as to whether it is considered that, given the initial identified risks, by applying appropriate measures or controls, the organisation will be able to bring the residual risk below the acceptable risk.

7. That all of the above information shall be recorded in the corresponding report, but:

a) It is essential that the report is clear, understandable and one of the best ways to do this is to make it graphic.

b) The measures contemplated in this report have to be integrated into a cycle of management and continuous improvement, which makes it advisable to use tools that allow such iteration and continuous improvement.

c) The report is one thing and the template to be submitted to the authority is another. Obviously this template conditions the minimum information and the format of the report, but in my opinion we cannot confuse, nor should the information to be reported (unless the template is very complete and demanding) coincide with the information to be collected and managed, which may be more, but not less than that of the template to be communicated.

d) The AIA does not speak of publication or communication to interested parties, but we understand that (except for aspects that for reasons such as "business" confidentiality or others that it is legitimate to ignore) it may be good practice to publish it and/or communicate it – at least – to interested parties.

8. It has not been the purpose of this chapter (as it has been addressed in other chapters) to analyse all the many possible intersections that may occur between AI systems and the processing of personal data, which are addressed in another chapter of this book, but to analyse the requirement set out in Article 27 of the AIA that if any of the obligations set out in that Article relating to FRIAs is already fulfilled by the data protection impact assessment under the GDPR, the FRIA assessment shall be carried out together with it, and we have concluded that it is one thing for them to be carried out – ideally – jointly; and another is that it is possible, and sometimes advisable, that without prejudice to the coordination of the two assessments, the data protection assessment should have its own *ad hoc* report, and in fact the AIA itself requires separate summaries for both: the DPIA on high-risk AI systems and the high-risk FRIA.

RISK MANAGEMENT SYSTEMS AS A SPECIFIC OBLIGATION FOR HIGH-RISK ARTIFICIAL INTELLIGENCE SYSTEMS IN ARTICLE 9 OF THE REGULATION

Pere Simón Castellano

Senior Lecturer in Constitutional Law
International University of La Rioja – UNIR[1]

I. What is a risk management system? Conceptual autonomy with respect to related figures

In this study we will deal with the regulation of risk management systems, which Article 9 of the AIA establishes as a legal obligation or a minimum essential requirement of any AI system that is classified as high risk.

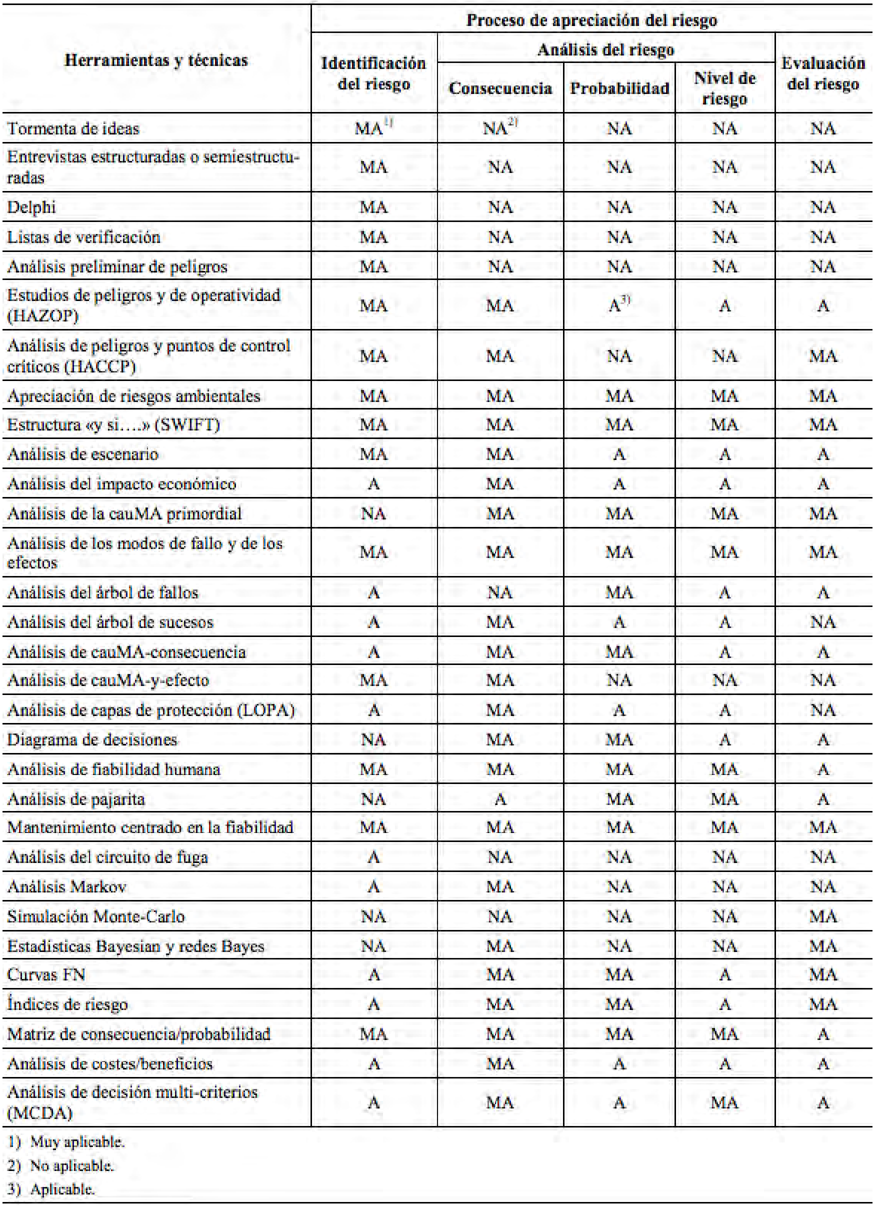
A risk management system is a set of processes, policies, procedures and tools designed to identify, assess, mitigate and monitor the risks faced by an organisation in achieving its objectives. These systems help organisations to understand and manage risks effectively in order to minimise losses, maximise opportunities and ensure business continuity.

Obviously, there are many types of management systems, depending on the particular sectoral area in which they apply and the risks they aim to minimise. Thus, risk management systems address risks of different types, including financial, operational, environmental, process quality, tax, information security, strategic, legal, privacy, generic compliance or criminal (criminal risk prevention) and reputational risks. These risks can arise from a variety of sources, such as market volatility, changes in the legal and regulatory environment, internal process failures, natural disasters, cyber-attacks, among others.

A risk management system generally follows a cyclical process that includes the following stages:

[1] This work is carried out in the framework of the Project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by MCIN/AEI/10.13039/501100011033/ ERDF, EU.

1. Risk identification (or assessment phase)



Herramientas y técnicas	Proceso de apreciación del riesgo				
	Identificación del riesgo	Análisis del riesgo			Evaluación del riesgo
		Consecuencia	Probabilidad	Nivel de riesgo	
Tormenta de ideas	MA[1]	NA[2]	NA	NA	NA
Entrevistas estructuradas o semiestructuradas	MA	NA	NA	NA	NA
Delphi	MA	NA	NA	NA	NA
Listas de verificación	MA	NA	NA	NA	NA
Análisis preliminar de peligros	MA	NA	NA	NA	NA
Estudios de peligros y de operatividad (HAZOP)	MA	MA	A[3]	A	A
Análisis de peligros y puntos de control críticos (HACCP)	MA	MA	NA	NA	MA
Apreciación de riesgos ambientales	MA	MA	MA	MA	MA
Estructura «y si….» (SWIFT)	MA	MA	MA	MA	MA
Análisis de escenario	MA	MA	A	A	A
Análisis del impacto económico	A	MA	A	A	A
Análisis de la cauMA primordial	NA	MA	MA	MA	MA
Análisis de los modos de fallo y de los efectos	MA	MA	MA	MA	MA
Análisis del árbol de fallos	A	NA	MA	A	A
Análisis del árbol de sucesos	A	MA	A	A	NA
Análisis de cauMA-consecuencia	A	MA	MA	A	A
Análisis de cauMA-y-efecto	MA	MA	NA	NA	NA
Análisis de capas de protección (LOPA)	A	MA	A	A	NA
Diagrama de decisiones	NA	MA	MA	A	A
Análisis de fiabilidad humana	MA	MA	MA	MA	A
Análisis de pajarita	NA	A	MA	MA	A
Mantenimiento centrado en la fiabilidad	MA	MA	MA	MA	MA
Análisis del circuito de fuga	A	NA	NA	NA	NA
Análisis Markov	A	MA	NA	NA	NA
Simulación Monte-Carlo	NA	NA	NA	NA	MA
Estadísticas Bayesian y redes Bayes	NA	MA	NA	NA	MA
Curvas FN	A	MA	MA	A	MA
Índices de riesgo	A	MA	MA	A	MA
Matriz de consecuencia/probabilidad	MA	MA	MA	MA	A
Análisis de costes/beneficios	A	MA	A	A	A
Análisis de decisión multi-criterios (MCDA)	A	MA	A	MA	A

1) Muy aplicable.
2) No aplicable.
3) Aplicable.

Figure 1. Detail of the degree of applicability of the tools used for risk assessment. Source UNE ISO 31010:2010

The assessment or identification phase consists of identifying and understanding the potential risks faced by the organisation in achieving its objectives. There are many techniques for this, which are shared below in table 1 derived from the ISO 31010:2010 risk assessment techniques. These are probably the most commonly used techniques, all of which are explained in particular detail in the international standard ISO 31010:2010, many of them exemplified with figures or diagrams. Logically, here we have only mentioned some of the techniques included in the international standard. Many of the techniques and tools described are shown in Figure 1, with details of their degree of application and effectiveness (1 – Highly applicable; 2 – Not applicable; 3 – Applicable) in the different phases of risk management.

2. Rights risk assessment and its differentiation from provider risk analysis and risk management

This phase involves analysing and assessing the probability of occurrence and the impact of the identified risks in order to prioritise them according to their importance. Assessing a risk involves considering all possible scenarios in which the risk would become effective. Risk assessment consists of evaluating the impact of exposure to the threat, together with the probability of the threat materialising. The impact, on the other hand, is determined on the basis of the possible damage that may occur if the threat materialises, e.g., an impact would be negligible if it had no consequences on the protected legal assets or, on the contrary, an impact would be significant if the damage caused to the protected legal assets would be critical. Depending on the probability and impact associated with the threats, it is possible to determine the level of inherent risk.

Risk assessment is inextricably linked to the risk matrix that is constructed depending on the method used. As regards methods for quantifying, examples and models of risk matrices and risk mapping, we refer to specific works on compliance systems[2].

The most commonly used risk matrices are 3x3 and 5x5 and usually include the factors of probability and impact or severity, although different matrices and formulas can be used that also apply other elements or criteria such

[2] See the works of Simón Castellano, P. "Responsabilidad penal de las personas jurídicas, mapa de riesgos y cumplimientos en la empresa", in Simón Castellano, P. y Abadías Selma, A. (coordinadores), *Mapa de riesgos penales y prevención del delito en la empresa*, Wolters Kluwer – Bosch, 2020, pp. 31-76 and Salvador Lafuente, A. "Mapa de riesgos: identificación y análisis de riesgos y controles", in Simón Castellano, P. and Abadías Selma, A. (coordinators), *Mapa de riesgos penales y prevención del delito en la empresa*, Wolters Kluwer – Bosch, 2020, pp. 78-119.

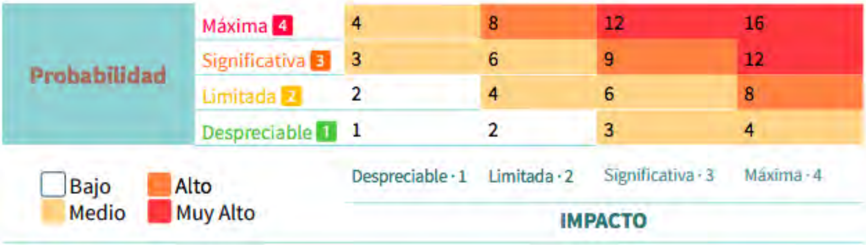
as function, substitution, depth, degree of externalisation, level of aggression and vulnerability.

The risk assessment phase within a risk management system must necessarily be differentiated from impact assessments in certain specific areas or also known as *ad hoc* risk assessments, such as data protection impact assessments and algorithmic impact assessments on fundamental rights. The latter will be the subject of study in this same work, *infra*, in the chapter headed by Eduard Chaveli, in relation to the content of Article 27 of the AIA, linked to algorithmic impact assessments on fundamental rights.

It is particularly useful here to draw an analogy with information security management systems (hereinafter ISMS), which are especially useful for ensuring the privacy and protection of personal data, which are unique and essential assets (information and data) of companies and public administrations. In this regard, there are a number of standards.

The AEPD has drawn up various guides for carrying out risk analysis within information security and data protection risk management systems, with the aim of establishing a roadmap to address the risks of personal data processing by establishing security measures and controls that guarantee the rights and freedoms of individuals in the field of privacy and data protection. It is in this area that we find the practical guide to risk analysis in the processing of personal data subject to the GDPR, whose approach is a mix of many of the principles and guidelines of ISO methodologies and management systems.

A good example is the risk matrix and the proposed formula for calculating inherent risk and residual risk, which is shared by the two guides and is well illustrated in Figure 2. The main advantage is that it is a very simple formula; the main disadvantage is that it loses level of risk detail relative to other formulas, which may include five or more levels of probability and impact.



Probabilidad	Máxima 4	4	8	12	16
	Significativa 3	3	6	9	12
	Limitada 2	2	4	6	8
	Despreciable 1	1	2	3	4
		Despreciable · 1	Limitada · 2	Significativa · 3	Máxima · 4

IMPACTO

□ Bajo 🟨 Medio 🟧 Alto 🟥 Muy Alto

Figure 2. Risk matrix. Source: AEPD Practical Guide

For the purposes of this paper, the analysis is structured in three phases based on the principle of proactive responsibility, communication, review and continuous improvement. These three phases are risk identification, risk analysis and risk treatment. This is illustrated in Figure 3.

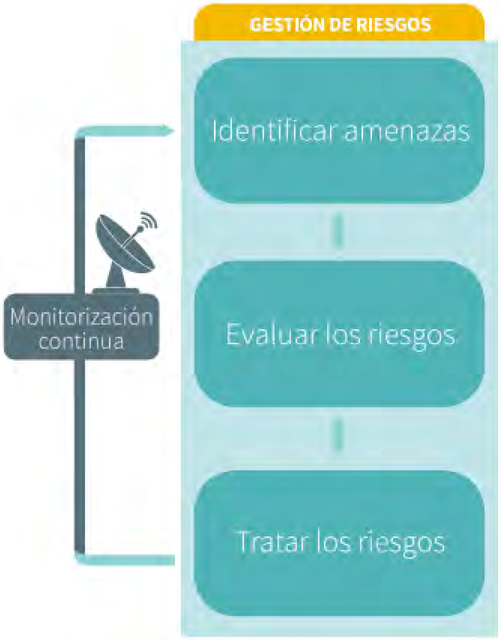


Figure 3. Three phases guided by the principle of monitoring or continuous improvement. Source: AEPD Practical Guide

Another good example of the above in relation to ISMS can be found in the international standards for the implementation and management of a management system. UNE-EN ISO/IEC 27001:2023, which is the European standard that in turn adopts the international standard on requirements for information systems management systems International Standard ISO/IEC 27001:2022, and which is complemented by ISO/IEC 27005, which provides guidelines for information security risk management.

This international standard provides guidance on the application of a process-oriented risk management approach to assist in the successful implementation of, and compliance with, the security risk management requirements of ISO/IEC 27001.

Let's take a look at a chart showing the detailed relationships between the ISO/IEC standards of the ISMS family.

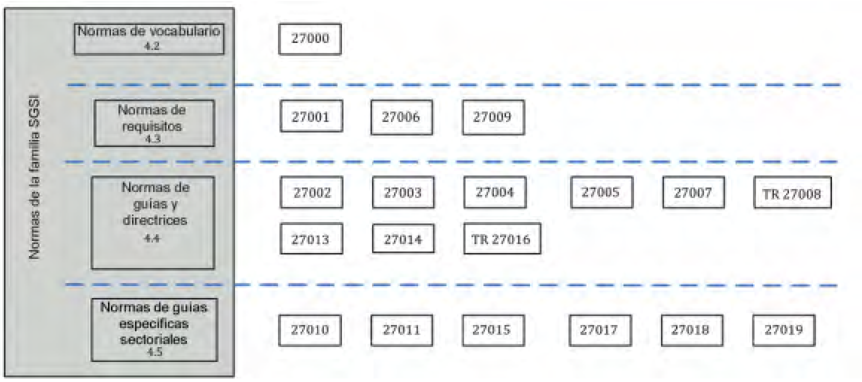


Figure 4. ISMS family standards. Source: UNE-EN ISO/IEC 27000:2019

ISO/IEC 27005 norm contains a number of general recommendations and guidelines for risk management in ISMS. It defines risk as a threat that exploits the vulnerability of an asset and may cause damage and relates risk to the use, ownership, operation, distribution, and adoption of enterprise information technology. The international standard uses a structured, systematic and rigorous risk analysis process for the creation of the risk treatment plan. This management system identifies the information assets to be protected, including personal data protection, and assesses the risks from the perspective of weaknesses or vulnerabilities and threats to which they are exposed, proposing controls to address the risk by reducing, accepting, transferring or even eliminating it.

The international standard is very detailed and has specific sections on risk matrix, defining the scope and limits of the security system, identifying and rating assets based on their impact, and quantifying the likelihood and impact of risk. It also suggests ways to evaluate vulnerabilities and traditional threats, and definitions of acceptable risk and criteria for its modification.

Figure 5 details the step from inherent risk to residual risk, in the treatment of acceptable risks as a result of satisfactory assessment.

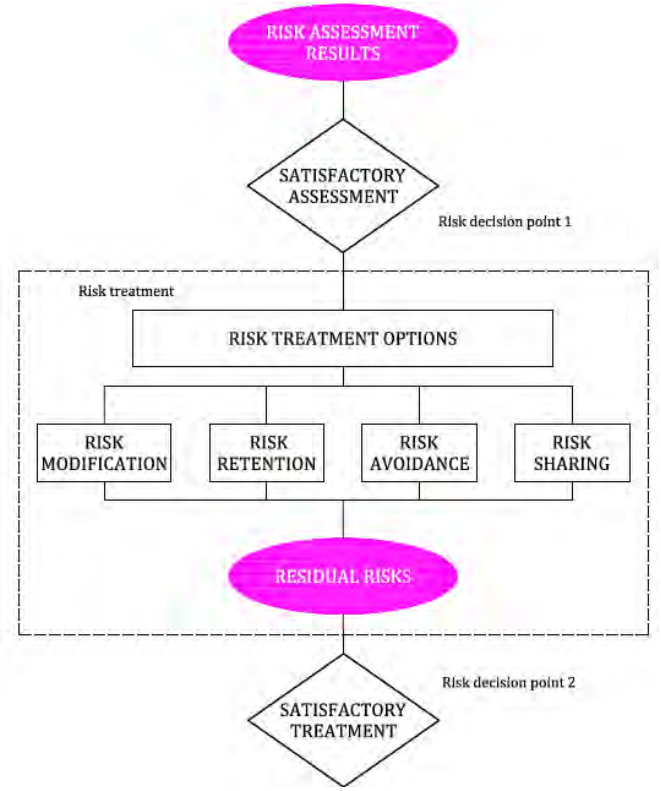


Figure 5. From assessment to acceptable treatment. Source: UNE-EN ISO/IEC 27005:2018

Be that as it may, the most recent international norm or standard, the aforementioned UNE-EN ISO/IEC 27001:2023, tells us that the organisation must develop and apply a process for assessing information security risks. This process must meet the following requirements: the organisation must define and maintain criteria for information security risks, including risk acceptance criteria and criteria for carrying out risk assessments.

It must also be ensured that successive information security risk assessments generate consistent, valid, and comparable results.

Regarding the identification of information security risks, this is achieved by conducting the risk assessment process to identify the risks associated with the loss of confidentiality, integrity and availability of information within the scope of the information security management system. It is also important to identify the owners of the risks.

In order to analyse information security risks, it is necessary to assess the potential consequences that would arise if the identified risks were to materialise, to make a realistic assessment of the likelihood of occurrence of the identified risks and to determine risk levels.

Finally, information security risks should be assessed by comparing the results of the risk analysis with the established risk criteria and prioritising the treatment of the analysed risks. In addition, the organisation should keep documented information on the information security risk assessment process.

As we can see, the risk assessment phase of a risk management system has nothing to do with what an impact assessment in that particular sectoral area is or implies; the equivalent of this would be a data protection impact assessment.

On this matter, to continue with the analogy, there is also a specific guide from the AEPD for carrying out an impact assessment, the methodology of which is different from the generic risk assessments of ISMS. The impact assessment (hereinafter, DPIA) is, above all, a tool for developing privacy by design within organisations, in the same way as the design and architecture of the register of processing activities and other risk assessments. The DPIA, like any other assessment, should be carried out for each processing activity, without prejudice to the possibility of extracting global indicators or grouped by business processes or departments.

The major difference between a DPIA and a standard risk assessment is primarily the data subjects' rights approach and the use of data protection principles as conceptual frameworks for the analysis of the risk of processing[3]. Moreover, the assessment focuses on a specific processing of personal data and its lifecycle (data and processing). The focus of the DPIA is on identifying threats to the rights and freedoms of the data subject, in a context of personal data processing, so that the DPIA, in short, is not a functional analysis of an information system in which technological risks are assessed, nor is it an information security audit or a compliance audit in general.

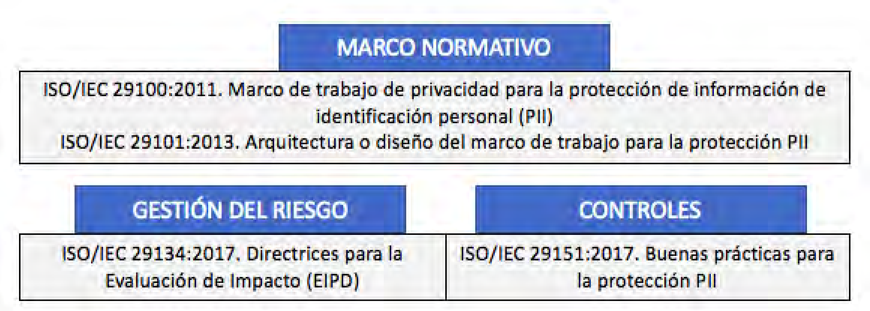
As a result of this approach the risk scenarios we will work with in an DPIA will be discrimination, identity theft, fraud, financial loss, reputational damage, loss of confidentiality of data subject to professional secrecy, unauthorised reversal of pseudonymisation, loss of control over personal data, disclosure of racial or ethnic origin of the data subject, disclosure of political opinion, religious or philosophical belief or trade union activism, disclosure of details about the health or sexual history of the data subject, disclosure of

[3] In this regard, see Simón Castellano, P. "El ejercicio de las funciones del delegado de protección de datos en la supervisión y gestión de procesos críticos", in Simón Castellano, P. and Bacaria Martrus, J. (coordinators), *Las funciones del delegado de protección de datos en los distintos sectores de actividad*, Wolters Kluwer – Bosch, 2020, pp. 27-74.

criminal convictions or administrative offences of the data subject, among others. If we look at the topics and approach, they are broader than in a generic risk assessment of a risk management system, despite the fact that this study is carried out in detail on a single processing of personal data or on a single data processing operation, taking into account the entire life cycle of the data.

As previously mentioned, the DPIA analyzes the privacy risks associated with the personal data processing activities and information systems involved in the organization. To this end, a specific international standard or norm has been developed, ISO/IEC 29134:2017, which incorporates the phases in which a DPIA must be carried out and the structure that the report resulting from the process must follow.

Beyond Article 35 of the GDPR, and the brief mention in Article 28.1 of the LOPDGDD, guidelines, methodologies, and controls for conducting a DPIA can be found in international standards, as indicated in the figure below.



MARCO NORMATIVO

ISO/IEC 29100:2011. Marco de trabajo de privacidad para la protección de información de identificación personal (PII)
ISO/IEC 29101:2013. Arquitectura o diseño del marco de trabajo para la protección PII

GESTIÓN DEL RIESGO

ISO/IEC 29134:2017. Directrices para la Evaluación de Impacto (EIPD)

CONTROLES

ISO/IEC 29151:2017. Buenas prácticas para la protección PII

Figure 6. DPIA framework, risk management and controls. Source: own elaboration

The DPIA can be carried out using different methods and tools. Neither the European nor the Spanish standard establishes a preference or obligation for a particular methodology or system. In any case, we have at our disposal the regulatory framework described in Figure 6, which includes ISO/IEC 29134:2017, specifically for data protection impact assessments. We also have at our disposal the AEPD's practical guide for data protection impact assessments and data protection impact assessments subject to the GDPR, which we have already mentioned above, although it is much less detailed or less detailed than the regulatory framework, risk management and controls provided in the international norms and standards mentioned in the illustration.

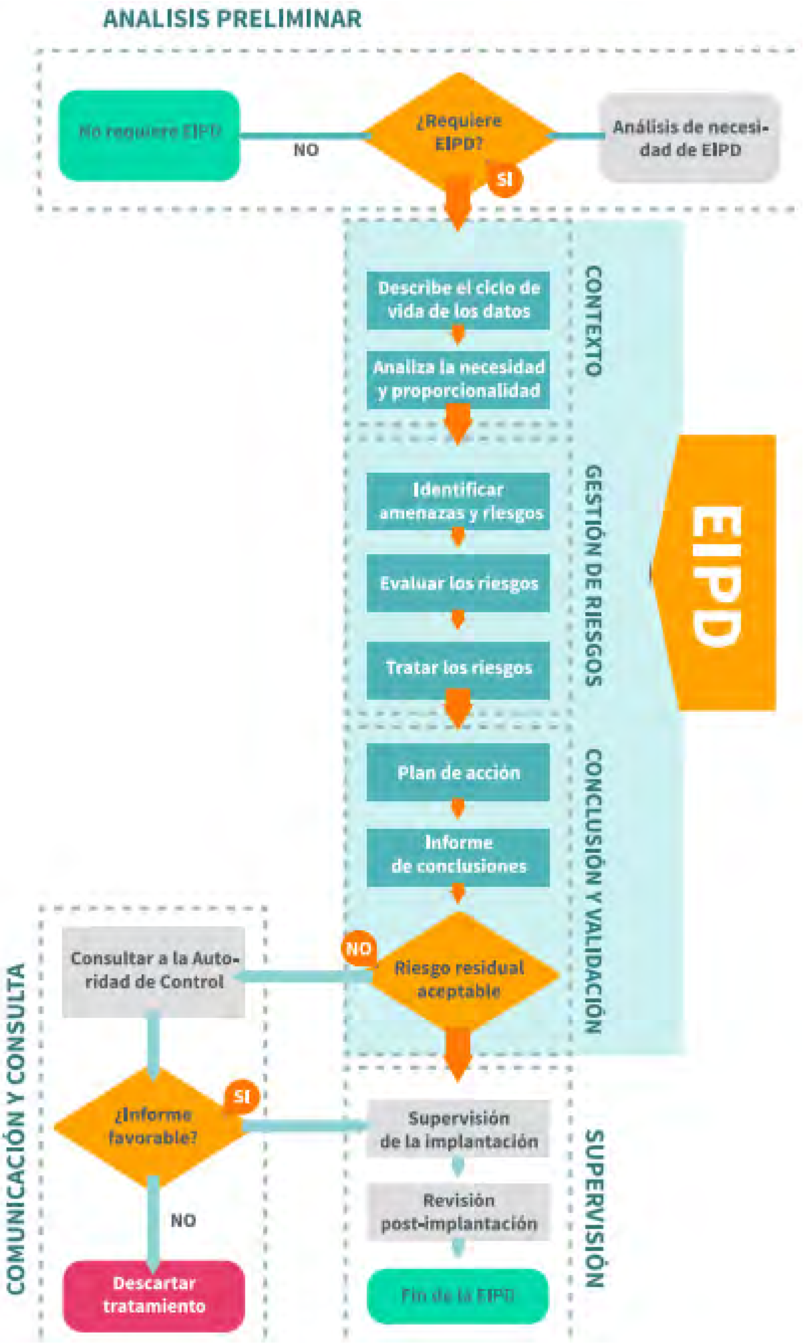


Figure 7. DPIA methodology. Source: AEPD DPIA Practical Guide.

However, there is an even more important difference between the obligations under Article 9 and Article 27 of the AIA, which have less to do with what or the content of a risk management system versus the subject of an algorithmic impact assessment on fundamental rights, and more to do with who is obliged to maintain the system or carry out the assessment. The former, Article 9, refers to a minimum requirement, and as will be seen below in Section II.2 by means of a summary table of obligations in relation to obliged subjects, it mostly projects effects on the providers of high-risk AI systems. On the other hand, Article 27 of the AIA focuses on and refers exclusively to users (this term was used at the draft stage) or "deployers" (final terminology in the last version known on 13 March 2024 with the approval of the final amendments by the European Parliament), which are the companies or public administrations that decide to use, employ or implement a high-risk AI system, regardless of the obligations of the manufacturer or provider, importers and distributors of such systems. Thus, the specific obligation to carry out an assessment of the algorithmic impact on fundamental rights falls exclusively on the users or those responsible for the deployment of the technological tool classified as high-risk.

3. Risk mitigation, monitoring and review

In this phase the main objective is to develop and implement strategies and measures to reduce the likelihood of occurrence or impact of risks, as well as to prepare to respond effectively in the event of their occurrence. It also includes continuous monitoring of the risks and mitigation measures implemented, as well as periodic review of the risk management system to ensure its effectiveness and relevance.

The final stage of the risk management process involves treating the risks to reduce or mitigate their effects. The objective of treating risks is to reduce their exposure level by implementing control measures to reduce the likelihood and impact, severity, or seriousness of their occurrence. Inherent risk can be treated with the objective of reducing or mitigating it, depending on the measure adopted, until the residual risk is at a considered reasonable level. The residual risk shall be the result of reducing the inherent risk level based on the effectiveness of the active controls, which is calculated, inter alia, taking into account the vulnerability percentage of the active controls. Vulnerability can be calculated in different ways, as explained in comprehensive or specific risk mapping studies[4] (criminal).

[4] See Simón Castellano, P. and Abadías Selma, A. (coordinators), *Mapa de riesgos penales y*

Be that as it may, for the purpose of this paper and by way of introduction, it is essential to understand the practical consequences of the principle of continuous improvement, which requires the constant or periodic review of compliance systems and programmes.

Continuous improvement is basically achieved through the regular involvement and encouragement of the compliance body, whether single or collective, in two fundamental areas of any risk management system: (1) the management of the company's registers and inventories and (2) the communication, consultation, monitoring, and review of previous risk assessments.

The registers and inventories make it possible to relate and connect the context of the organisation with the risks and also with the controls derived from the risk assessments, thus providing an up-to-date view of the organisation's needs. It is a functional, operational, and actionable tool, which should be consulted and reviewed by the compliance officer on a recurring basis, or by the head of the management system. In fact, the compliance body or the compliance officer (or manager, since depending on the type of system, the names of the managers and/or positions may vary) should be informed of any changes or modifications to these inventories.

Continuous improvement is also achieved by commissioning internal or external audits, and by managing and resolving specific incidents or by identifying responsibilities through specific channels, such as whistleblowing channels, which must necessarily include measures to protect the whistleblower or claimant[5]. The continuous improvement process requires defining a plan of audits and periodic reviews based on the organisation's activities and business processes, as well as on the results of risk assessments.

A well-established and executed risk management system helps organisations to make informed decisions, improve their resilience to risks and create long-term value, but obviously only in the areas where it has or projects effects: environmental, quality, financial, criminal or regulatory compliance, information security, data protection or AI system security and resilience, among many other possibilities.

What in particular can risk management methodologies and systems contribute to the implementation of a high-risk AI system? Risk management approaches applied throughout the AI system lifecycle can identify, assess,

prevención del delito en la empresa, Wolters Kluwer – Bosch, 2020.

[5] See the works of León Alapont, J. *Canales de denuncia e investigaciones internas en el marco del compliance penal corporativo*, Tirant lo Blanch, 2023; Simón Castellano, P. "La inmunidad penal como recompensa a los denunciantes. Allende un nuevo factor subjetivo-formal de punibilidad", *Revista electrónica de ciencia penal y criminología,* n.º 24, 2022.

prioritise, and resolve situations that could adversely affect a system's performance and outcomes.

Different phases can be identified to manage AI risks while ensuring respect for human rights and democratic values, without confusing, as mentioned above, this internal governance and management process with what is involved in conducting a fundamental rights impact assessment. Risk management systems can be based on the NIST AI Risk Management Framework mentioned above, the ISO 31000 family risk management framework, which we have also detailed above, and the Organisation for Economic Co-operation and Development (OECD) due diligence guidance[6].

These different phases could be classified, following the above-mentioned OECD guidance, as follows: (1) defining the scope, context, and criteria, including relevant AI principles, stakeholders, and actors for each phase of the AI system lifecycle and for the lifecycle itself; (2) risk assessment phase for trusted AI by identifying and analysing problems at individual, aggregate, and societal levels, and assessing the likelihood and level of harm (e.g., small risks may accumulate and become a larger risk); (3) treating risks to cease, prevent, or mitigate adverse impacts, commensurate with the likelihood and extent of each; (4) governing the risk management process by embedding and cultivating a risk management culture in organisations; monitoring and reviewing the process on an ongoing basis; and documenting, communicating, and consulting on the process and its outcomes.

The only way to achieve reliable and accountable AI is for the actors involved to take advantage of processes, indicators, standards, certification schemes, audits, and other mechanisms to monitor and guarantee these processes and components at each stage of the AI system's life cycle. This should be an iterative process where the findings and results of one stage of risk management feed into the others, achieving a kind of continuous improvement scenario. And it is in this sense that it is easy to identify the differences between a risk management system and an algorithmic impact assessment on fundamental rights, which has an exclusive focus on risk and its derivatives, a greater focus in terms of the potential impact (groups affected, rights and principles affected, duration in time, proportionality, etc.) but much more limited in terms of a specific treatment, processing or use of the AI technology.

Artificial Intelligence risks can be assessed at different levels, including at the governance and process level, focusing on risks related to value-based

[6] See OECD (2018), *OECD Due Diligence Guidance for Responsible Business Conduct*, available at http://mneguidelines.oecd.org/OECD-Due-Diligence-Guidance-for-Responsible-BusinessConduct.pdf (last accessed 9 March 2024).

principles (e.g., accountability), and at the technical level, focusing on technical risks (e.g., robustness and performance), and underlying sub-risks (e.g., statistical accuracy).

One step toward ensuring accountability in AI is to link principles, rights, and risks to specific procedural and technical attributes. While some existing frameworks (models of risk management systems in AI environments) provide AI actors with substantial guidance, such as the taxonomy of AI trustworthiness in Newman[7] (2023) or the taxonomy of AI legal safeguards in Simon[8] (2023), turning value-based principles into specific technical requirements and attributes is an evolving field, and in any case cannot be exhaustive insofar as there is no ideal management model, but as many models as there are companies and administrations with business processes, contexts, data processing, nature and scope of AI and its singular or unique use.

II. Evolution of the meaning, content, and recipients of the obligation to have a risk management system (Article 9 AIA)

The regulations under analysis in this chapter, and more specifically the legal obligation to design, implement and monitor a risk management system, are set out in Chapter III of the AIA, entitled "High-risk AI systems", which contains the rules for the classification of high-risk AI systems in its first section, while the second section, where the obligation in Article 9 of the AIA is located, sets out the mandatory minimum requirements for high-risk AI systems.

These requirements are, in turn, a derivative of the ethical guidelines for trustworthy Artificial Intelligence that were developed by the independent high-level expert group on Artificial Intelligence set up by the European Commission in June 2018[9]. Adaptability is considered in relation to the technical solutions required to achieve compliance with the above requirements, which may derive from regulations or technical specifications, or be developed according to specific scientific or sectoral knowledge.

[7] Newman, J., "A Taxonomy of Trustworthiness for Artificial Intelligence: Connecting Properties of Trustworthiness with Risk Management and the Lifecycle", *UC Berkeley*, 2023, available at https://cltc.berkeley.edu/wpcontent/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf (last accessed 10 March 2024).

[8] Simón Castellano, P. "Taxonomía de las garantías jurídicas en el empleo de los sistemas de inteligencia artificial", *Revista de Derecho Político,* n.º 117, 2023, pp. 153-196.

[9] European Union, *Ethical Guidelines for Trustworthy AI. High Level Expert Group on Artificial Intelligence*, European Commission, Brussels 2019.

In this respect, a wide margin of discretion is granted to the AI system provider to determine how to meet the requirements, taking into account the current state of technology and scientific and technological developments. Thus, we are dealing with mandatory minimum requirements that can be achieved in different ways: within the scope of Article 9 of the AIA, one can opt for models uniquely designed within the context, scope and nature of the organisation and the AI to be used; or follow the requirements of certain international standards such as ISO/IEC 42001:2023, ISO/IEC TR 24030:2021 and ISO/IEC TR 5469:2024 (which are briefly summarised, among other standards that help in the interpretation of these three standards – e.g., 24027:2023 dealing with biases or 22989:2022 dealing with concepts and terminology – in the chapter below by Eduard Chaveli), or the NIST 800 218 PW.I.I.; NIST 800 218RV.1.i.; or the OECD's *High-level AI risk-management interoperability framework*.

ISO 42001 aims to assist organisations in responsibly fulfilling their roles in relation to AI systems, including their use, development, monitoring, and provision of products or services. AI raises specific considerations such as the use of AI for automatic or automated decision making, sometimes in a non-transparent and non-explanatory way, may require specific handling beyond the handling of classical IT systems; the use of data analytics, perception and machine learning, rather than human-coded logic to design systems, which increases the application opportunities for AI systems and changes the way such systems are developed, justified and deployed; AI systems that perform continuous learning and change their behaviour during use, which requires special consideration, in line with the continuous improvement and iterative nature of management systems, to ensure that their responsible use continues in the face of changing behaviour.

ISO 42001 provides requirements for establishing, implementing, maintaining and continually improving an AI management system in the context of an organisation. Organisations are expected to focus their application of requirements on characteristics that are unique to AI. Certain characteristics of AI, such as the ability to learn and continuously improve or the lack of transparency or explainability, may warrant different safeguards if they raise additional concerns compared to how the task would traditionally be performed.

The adoption of an AI management system to extend existing management structures is a strategic decision for an organisation. The organisation's needs and objectives, processes, size and structure, as well as the expectations of the various stakeholders, influence the establishment and implementation of the AI management system. Another set of factors influencing the establishment and implementation of the AI management system are the many use cases for AI and the need to find the right balance between governance mechanisms and innovation. Organisations may choose to implement these requirements using a risk-based approach to ensure that the appropriate level of control is applied for the particular AI use cases, services or products within the scope of the organisation. All of these influencing factors are expected to change and be reviewed from time to time.

The AI management system should be integrated with the organisation's processes and overall management structure. Specific AI-related issues should be considered in the design of processes, information systems and controls. The model proposed by ISO 42001 sets out a set of guidelines for the implementation of applicable controls to support these processes, and avoids specific guidance on management processes. The organisation can combine generally accepted frameworks, other international standards and its own experience to implement critical processes such as risk management, lifecycle management, and data quality management that are appropriate to the specific AI use cases, products, or services within scope.

The ISO 42001 standard itself indicates that an organisation that complies with the requirements is one that can generate evidence of its responsibility and accountability for its role in relation to AI systems. This standard applies a harmonised structure, with identical clause numbers, clause headings, common text and terms, and core definitions, which is developed to improve alignment between management system standards and make it compatible with other international AI risk management system standards. The AI management system provides specific requirements for managing the issues and risks arising from the use of AI in an organisation. This common approach facilitates implementation and consistency with other management system standards, e.g., related to quality, safety, security and privacy.

Another good example, as mentioned above, is the OECD's *High-level AI risk-management interoperability framework* for AI risk management. Figure 8 shows the structure that the OECD proposes for the design and minimum components (principles, AI system lifecycle and risk management phases) of a risk management system in the context of using an AI-based system; in Figure 9, on the other hand, we can see the components through a functional view that highlights the importance of communication and consultation, documentary evidence and monitoring, and review processes to achieve continuous improvement throughout the entire lifecycle of the AI system.

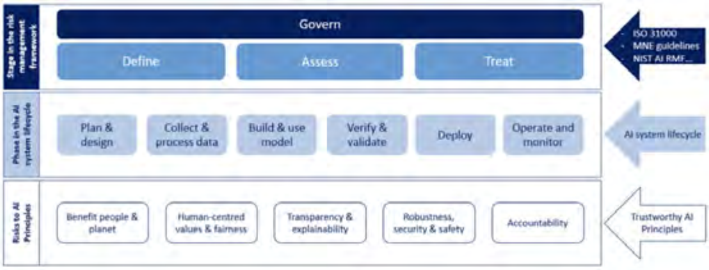


Figure 8. Structure of a risk management and interoperability system for a high-risk AI system. Source: OECD report entitled "Advancing accountability in AI" available at https://doi.org/10.1787/2448f04b-en.

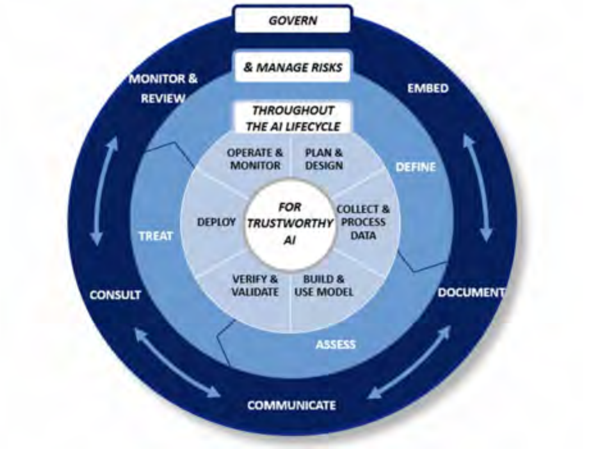


Figure 9. Functional view of a risk management and interoperability system for a high-risk AI system. Source: OECD report entitled "Advancing accountability in AI" available at https://doi.org/10.1787/2448f04b-en.

Governing the risk management process is key to achieving reliable Artificial Intelligence. Governance is a cross-cutting activity consisting of two main elements. The first element concerns the governance of the risk management process itself and includes oversight and review, documentation, communication, and consultation on the process and its outcomes. The second element of governance ensures the effectiveness of the risk management process by embedding it in the wider governance culture and processes of organisations.

In any case, the minimum requirements of Chapter III of the AIA for high-risk systems, which include at the top the significant and mandatory mention of risk management systems, are a set of horizontal obligations imposed on providers of high-risk AI systems[10], although Chapter III of

[10] On this issue see Simón Castellano, P. *Justicia cautelar e inteligencia artificial: la alternativa a los atávicos heurísticos judiciales*, J. M. Bosch, 2021; Cotino Hueso, L., "Los usos de la inteligencia

the AIA also sets out a series of minimum requirements or obligations for users (understood as users in the sense of the AIA, i.e., any company or organisation that uses or employs AI systems; in no way equating "users" with end-users or recipients resulting from the application of AI systems) and other agents or actors such as importers, distributors, and authorised representatives[11]. This is intended to strengthen the effectiveness of existing rights and remedies by establishing specific requirements and obligations, in particular on transparency, technical documentation and registration of AI systems. Requirements should apply to high-risk AI systems in terms of risk management, quality, and relevance of data sets used, technical documentation and record keeping, transparency and provision of information to users, human oversight, robustness, accuracy and cybersecurity. These requirements are necessary to effectively mitigate the risks that the use of AI potentially projects for legal goods as diverse as privacy, security, health, equality, or freedom of information, among many others.

1. What is a risk management system according to the Regulation? The content of the obligation

The European legislator has made an enormous effort to realise the specific obligation or requirement contained in article 9 of the AIA, and has done so through the recitals, the main ideas of which we will try to bring together and systematise in the following lines. An AI risk management system is a set of measures designed to identify, assess, and mitigate the risks associated with AI systems considered to be high risk, which are placed on the market or put into service. The objective is to ensure a high level of reliability and accountability of high-risk systems by applying certain minimum and prescriptive requirements, taking into account the intended purpose and context of use of the AI system. Measures taken by providers to comply with the mandatory requirements of the AIA should take into account the generally recognised state of the art in Artificial Intelligence, be proportionate, and effective in meeting the objectives of the AIA (see Recital 42 of the AIA).

artificial en el sector público, su variable impacto y categorización jurídica", *Revista Canaria de Administración Pública*, n.º 1, 2023, pp. 211-242; Presno Linera, M. A. *Derechos fundamentales e inteligencia artificial*, Marcial Pons, 2022; Presno Linera, M. A., "La propuesta de "Ley de Inteligencia Artificial" europea", *Revista de las Cortes Generales*, n.º 116, 2023, pp. 81-133.

[11] In this regard, we recommend the work of Ramón Fernández, F., "Inteligencia artificial y transparencia en relación con la regulación de los servicios y mercados digitales", in Cobas Cobiella, M. E. and Guillén Catalán, R. (eds.), *Equidad y transparencia en la prestación de servicios*, Dykinson, Madrid, 2023, pp. 147-169.

Following the approach of the New Legislative Framework and the EU Digital Strategy, as set out in the Commission's Blue Guide Notice on the implementation of the EU product standards 2022 (C/2022/3637), the general rule is that several pieces of EU legislation may have to be taken into account for a product, as making available or putting into service can only take place when the product complies with all applicable EU harmonisation legislation. The hazards of AI systems covered by the AIA requirements concern different aspects than the existing Union harmonisation provisions. This requires a simultaneous and complementary application of the various pieces of legislation.

In order to ensure consistency and avoid unnecessary administrative burdens or costs, providers of a product containing one or more high-risk AI systems should have flexibility in operational decisions on how to ensure compliance of a product containing one or more AI systems with all applicable requirements of Union harmonisation legislation in the best possible way. This flexibility should in no way undermine the provider's obligation to comply with all applicable requirements, including, significantly, the requirement to have an operational risk management system in place.

Recital 42a) of the AIA states that the risk management system should consist of a continuous and iterative process that is planned and executed throughout the life cycle of a high-risk AI system. The idea of continuous improvement and technology in motion requires the design of a management system that, in any case, must evolve as the technological context and scope, its specific uses within the organisation, and its projected effects change. This iterative, non-static process must aim to identify, assess and mitigate the relevant risks of Artificial Intelligence systems to health and safety (liability for material damage or defective products) and also to the fundamental rights of individuals, albeit in a generic and not detailed manner because that is the subject of another obligation, the one foreseen *ex art.* 27 AIA, with the assessment of algorithmic impact on fundamental rights.

The risk management system should be regularly reviewed and updated to ensure its ongoing effectiveness, as well as the justification and documentation of any significant decisions and actions taken under the AIA (again, we follow Recital 42a). This process should ensure that the provider identifies risks or adverse impacts and implements mitigation measures for known and reasonably foreseeable risks of AI systems to health, safety and fundamental rights, taking into account their intended purpose and reasonably foreseeable use, including potential risks arising from the interaction between the AI system and the environment in which it operates.

In the treatment phase, the risk management system should adopt the most appropriate risk management measures in light of the state of the art in

Artificial Intelligence. In identifying the most appropriate risk management measures, the provider should document and explain the decisions taken and, where relevant, involve external experts and stakeholders. When identifying reasonably foreseeable use of high-risk AI systems, the provider should cover uses of AI systems that, although not directly covered by the intended purpose and specified in the instructions for use, can be easily expected as a result of readily foreseeable human behaviour in the context of the specific characteristics and use of the particular AI system. Any known or foreseeable circumstances related to the use of the high-risk AI system in accordance with its intended purpose or under reasonably foreseeable conditions of use, which may give rise to risks to health, safety, or fundamental rights, should be included in the instructions for use provided by the provider.

Part of what goes into the management system also requires transparency. The purpose is to ensure that the user is aware of and takes into account those foreseeable risks when using the high-risk AI system. Identifying and implementing risk mitigation measures for foreseeable uses under the AIA should not require additional training measures specific to the high-risk AI system by the provider to address them. However, the AIA encourages providers to consider such additional training measures to mitigate reasonably foreseeable uses as necessary and appropriate.

The final wording of Article 9 of the AIA provides a clear delineation of what is meant by "risk management system" in relation to high-risk AI systems. It indicates that the management system should be understood as a continuous iterative process, planned, and executed throughout the lifecycle of the AI-based technology system and requiring regular systematic reviews and updates. As a minimum, Article 9.2 of the AIA requires it to consist of the following phases or stages:

> *"(a) the identification and analysis of the known and the reasonably foreseeable risks that the high-risk AI system can pose to health, safety or fundamental rights when the high-risk AI system is used in accordance with its intended purpose;*
> *(b) the estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose, and under conditions of reasonably foreseeable misuse;*
> *(c) the evaluation of other risks possibly arising, based on the analysis of data gathered from the post-market monitoring system referred to in Article 72;*
> *(d) the adoption of appropriate and targeted risk management measures designed to address the risks identified pursuant to point (a).*

Article 9.3 of the AIA states that the risks referred to in this Article concern only those which can be reasonably mitigated or eliminated through the development or design of the high-risk AI system, or the provision of adequate technical information. This directly links the management system to the system development or design phase (technical or organisational, both in the code development phase and potential bias errors in the data mining and training phase) and the phase of provision of adequate technical information (transparency and the rights of end users to know what safeguards and rationale are operating behind the AI tool). This paragraph could also be interpreted in the sense of differentiating the risks from those that must be thoroughly and exhaustively assessed in the context of the obligation under Article 27 of the AIA to carry out an algorithmic impact assessment on fundamental rights.

Article 9.9 of the AIA makes special mention of the need for guardianship and protection of minors, as well as other vulnerable groups. This is an indeterminate legal concept that the European AI Office and the Spanish AESIA must clarify and delimit.This means that when implementing any risk management system in the field of high-risk AI systems, providers must pay special attention to any potential negative effects on individuals under the age of eighteen and other vulnerable groups. They must establish ad hoc measures or controls to mitigate these risks.

Very interesting and relevant for practical purposes is the content of Article 9.10 of the AIA, which states that "For providers of high-risk AI systems that are subject to requirements regarding internal risk management processes under other relevant provisions of Union law, the aspects provided in paragraphs 1 to 9 may be part of, or combined with, the risk management procedures established pursuant to that law". This means that in case of co-existence with other management systems required by EU law, management systems may be coordinated, or even integrated, either by sectoral management systems (more difficult to imagine, but for example, integrations could be made with environmental and sustainability risk management systems or ethical and socially responsible management systems) or by specific information security systems (those required by ENISA and ENS or those specific to critical infrastructure).

On the other hand, Article 9.4 to 8 of the AIA focuses on risk management measures applicable to high-risk AI systems. On the one hand, it makes consideration of combined effects and interactions, i.e., the AIA stresses the importance of considering the effects and interactions resulting from the combined application of the requirements set out in the management system. The aim is to minimise risks more effectively while achieving an appropriate balance in the application of measures to meet those requirements. In other words, it seeks to strike a balance between effectiveness in risk management and fair application of the required measures. This is found in Article 9.4, where it refers to the effects and possible interaction arising from the combined application of requirements. In paragraph 5, on the other hand, we turn to the assessment and consideration of residual risks, where it is stated that risk management measures should take into account the relevant residual risks associated with each hazard, as well as the overall residual risk of high-risk AI systems. This implies that even after mitigation measures have been implemented, certain risks may persist, and it is important to assess and accept these residual risks appropriately. Where technically feasible, detection and assessment mechanisms are required to be established in the design and development of AI solutions, mitigation and control measures are required to be implemented, and training for controllers and deployers is required. High-risk AI systems shall be tested to determine the most appropriate and targeted risk management measures. Such testing shall verify that high-risk AI systems operate in a manner consistent with their intended purpose and meet the mandatory minimum requirements.

Ultimately, these AIA articles emphasise the importance of comprehensive risk management for high-risk AI systems, including consideration of the combined effects of measures and assessment of the residual risks associated with these systems. They also stress the need to design and develop AI systems in a way that minimises risks as far as possible by providing adequate information and training to those responsible for deployment. High-risk AI systems adversely affect persons under the age of eighteen and, where appropriate, other vulnerable groups of persons. And it opens the door to test environments, stating that testing of high-risk AI systems shall be carried out, as appropriate, at any time during the development process and in any case prior to their introduction to the market or putting into service. Testing shall be performed using pre-defined parameters and probability thresholds that are appropriate for the intended purpose of the high-risk AI system.

2. Obligated subjects. Who is obliged to have a risk management system? Summary table of obligations related to high-risk systems

In order to better understand the obligation to have a risk management system, it is necessary to understand the microcosm of agents or actors in the field of AIA and the different obligations and requirements that apply to each of them; what for some is an obligation or minimum requirement for others may be a requirement to verify that a third party has obtained a certification, has implemented a management system or has complied in due time and form with those derived from the AIA. Below is a summary table of the obligations for high risk systems, with those derived from or linked to the existence of a management system highlighted in bold, even if they may result or be interrelated with other variables.

High-risk AI systems Minimum requirements to be met by systems	● Providers of high-risk systems shall: ○ Establish, implement, document, and maintain a **Risk Management** system associated with the AI system, aiming to minimise risks to users and affected persons and demonstrating compliance with the requirements of current legislation, even after the products have been placed on the market. It shall pay particular attention to risks to health, safety, and fundamental rights. Establish a **Governance and Management Data** system for training and testing data, ensuring good practices in their design, collection, and preparation. In addition, they will have to ensure their relevance, correctness and appropriate statistical properties, avoiding biases that negatively affect individuals. ○ High Risk AI systems shall be accompanied by **updated Technical Documentation** demonstrating that the requirements are met before they are placed on the market, and throughout the time they are on the market. They shall automatically take **System Activity Logs ("logs")** throughout the life of the system. High Risk AI systems shall be **designed and developed in such a way as to ensure that their operation is sufficiently transparent** (special consideration shall be given in the design and development of the AI system in the framework of risk assessment and risk treatment, in particular where there are potential vulnerable groups or minors who may be end-users or recipients of such tools) to enable users to interpret the output of the system and to use such information appropriately. Information such as system capabilities, equipment requirements, scope of application, level of accuracy, human monitoring systems, etc., shall be provided. They shall allow High Risk AI systems to be **supervised by persons** during use to minimise risks to health, safety and fundamental rights, with particular attention to residual risks after implementation of mitigation measures. Users will be able to monitor the systems and interpret their output information. For real-time remote biometric identification, output will require separate verification and confirmation by at least two natural persons (with some exceptions contained in the law). Human oversight is an intrinsic element of the management systems, with definition of users, assignment of roles with different powers depending on their role in risk management, and treatment (in particular, controllers) and assurance through communication and consultation of the system. **Ensure an adequate level of accuracy, robustness and cybersecurity**, which will be declared in the accompanying technical documentation. In this respect, we refer to the chapter written by Professor Francisca Ramón Fernández in this same collective work. The management system implies supervising that the design of the AI tool is carried out with the maximum possible resistance to errors, biases, failures, or inconsistencies that may occur, especially in the interaction with other people or systems. In any case, they will incorporate cybersecurity measures appropriate and proportionate to their circumstances, with special attention to protection against manipulation of data training.

High-risk AI systems	As a consequence of the diversity of parties involved in the implementation, marketing, and operation of AI systems, especially high-risk ones, the AIA establishes differentiated obligations for each of them.
Obligations of providers, users and third parties	● **Providers**: providers are any natural or legal person, public authority, agency or other body that develops an AI system or a general-purpose AI model or that has an AI system or a general-purpose AI model developed and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge. They shall have the following obligations: ○ ensure that their AI systems comply with the requirements of the previous section (minimum requirements to be met by the systems), including the name or trademark and the address where it can be contacted. ○ Have a documented and updated quality management system, maintaining full documentation of the system (again, we refer to the contribution in this collective work by Professor Francisca Ramón Fernández). ○ They shall have custody of the system logs under their control. ○ ensure that the AI system is subject to an appropriate conformity assessment procedure before being placed on the market and/or put into service. ○ Cooperate with the authorities by recording the system, demonstrating compliance with all requirements of the Regulation when required to do so, and reporting non-compliances and risks identified and corrective actions taken as a result. ○ In the case of a provider established outside the EU, before placing their systems on the EU market, they must appoint by written mandate an authorised representative located in the EU. **Importers**: An importer is any natural or legal person located or established in the Union that places on the market an AI system that bears the name or trademark of a natural or legal person established in a third country. Before placing the system on the market they will have to ensure that it is in conformity with the regulation by verifying that: ○ The system provider has carried out the relevant conformity assessment procedure. ○ The provider has drawn up the necessary technical documentation. ○ The system is CE marked as required and is accompanied by the EU declaration of conformity and its instructions for use. ○ The provider has appointed an authorised representative in the EU.

	If any of these requirements are not met, or if there is sufficient reason to believe that such documentation is falsified or accompanied by forged documents, the importer shall refrain from placing the system on the market. ○ Importers shall cooperate with the competent authorities and inform them of their name or trademark on the product, together with the address where they can be contacted. ● **Distributors**: A distributor is any natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market without influencing its properties. Before making an AI system available on the market it must be ensured that: The system is CE marked as required and is accompanied by a copy of the EU Declaration of Conformity and its instructions for use. ○ That the provider and the importer have complied with the obligation to indicate the name or trademark and contact address and that the provider has a quality management system. If any of these requirements are not met, or if there is sufficient reason to believe that such documentation is falsified or accompanied by forged documents, the importer shall refrain from placing the system on the market. ○ Distributors shall cooperate with the competent authorities and inform them of their name or trademark on the product, together with the address where they can be contacted. ● **Users or deployers**: A user is any natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity. They shall have the following obligations: ○ Take appropriate technical and organisational measures to ensure that the use of such a system is in accordance with the accompanying instructions for use. ○ Exercise humane supervision of the system, ensuring that the person in charge has the necessary competence, training, authority and support. ○ Monitor the operation of the system. ○ Custody of system logs under its control. ○ Cooperate with the competent authorities. Users or those responsible for the deployment that are or belong to the public sector, as well as private operators that provide public services or those companies that assess credit and equity solvency and those that assess risks to set prices in health and life insurance, will have to carry out an additional assessment of the algorithmic impact on fundamental rights that the use of such a system may cause (we refer to the differentiation made in section I.2.1 of this work and to the chapter signed by Eduard Chaveli in this collective work).

Responsibilities in the value chain	Any distributor, importer, user (or deployer) or third party may be considered a provider of a high-risk AI system, and therefore be subject to the obligations required for providers under the Regulation in the following cases: ○ If they place their name or trademark on a High-Risk AI system already placed on the market or put into service, without prejudice to contractual arrangements stipulating that the obligations are otherwise assigned. ○ If it introduces a substantial modification to a High Risk AI system that has already been placed on the market or put into service, it will remain a High Risk system. It shall also apply in cases where the intended purpose of an AI system, including general-purpose AI systems, which have not been classified as High Risk and which have already been placed on the market or put into service, is changed in such a way that the AI system becomes a High Risk system. ● In the case of High Risk AI systems that are product safety components, the manufacturer shall be considered to be the provider of the AI system in the following cases: ○ The AI system is marketed together with the product under the name or trademark of the product manufacturer. ○ The AI system is put into service under the name or trademark of the product manufacturer. ○ The AI system is put into service under the name or trademark of the product manufacturer after the product has been placed on the market.

Figure 10. Summary table of obligations of high-risk systems and identification of obligated parties for each of these (obligations of providers, users/deployers, importers, distributors and third parties). Source: Font Advocats

III. Recapitulation and conclusions

In this paper, we looked at risk management systems as a minimum requirement for high-risk AI systems and how they relate to other specific obligations of providers. We did this by highlighting the important differences between this obligation and others. For example, some obligations only apply to providers, like the need to put the AI system through the relevant conformity assessment procedure or keep a quality management system and keep technical documentation safe (arts. 11, 17, and 18 of the AIA), while others only apply to "former" users and not to providers. In the latest version of the AIA, these "former" users are referred to as deployers, who carry out the mandatory obligation to conduct algorithmic impact assessments on fundamental rights.

In an attempt to synthesise as much as possible, we can state that the conclusions we have reached are as follows:

First. On the content of a risk management system. The risk management system is an iterative process of continuous improvement, planned and executed throughout the lifecycle of a high-risk AI system, which will require periodic systematic reviews and updates. It comprises at least three phases (identification, assessment, and treatment) and requires the incorporation of responsible persons and users with roles and functions within the system, in particular to monitor residual risk levels and the effectiveness and vulnerability of controls. As it is an obligation that is projected onto the providers or manufacturers, having a management system is a requirement that starts from the very design of the technological solution and does not end with the first marketing, but also requires post-market surveillance (in relation to Article 72 of the AIA). At a minimum, the management system should identify known and foreseeable risks to health, safety, or fundamental rights that the high-risk AI system may entail, as well as estimate and assess the levels of risk in relation to the possible intended uses and purposes. This analysis and assessment shall identify needs and vulnerabilities, which shall be addressed through the adoption of appropriate and specific risk management controls and measures designed to address the risks identified in advance.

Second. The need to differentiate this obligation from the so-called algorithmic impact assessment on fundamental rights (Article 27 AIA). There is an important substantive difference. The management system is made up of three phases and the algorithmic impact assessment on fundamental rights should be framed practically and exclusively in the assessment phase, since if the result is that the deployment of AI poses an unacceptable risk from the point of view of fundamental rights then there is no control to apply and no possible treatment phase. Only an impact assessment with a favourable outcome will allow the deployment by the user or deployer of that AI, classified as high-risk, in the specific business or public sphere (context, scope and nature) and for the specific purpose of the scenario. Having a management system in place from the inception or design of a technology, including the entire lifecycle of the AI system, is not the same as carrying out an algorithmic impact assessment on fundamental rights for a specific use in a specific business context. The obligations,whose content is different, also have different recipients or obliged parties. The former concerns providers (and distributors and importers to the extent that they must check and verify that the provider has an operational risk management system in place), whereas the algorithmic fundamental rights impact assessment concerns only "users" or "deployers". Moreover, it is an obligation that only arises or applies to certain deployers, more specifically, it only applies to those who are bodies governed by public law, or to private entities providing public services, or to companies assessing creditworthiness and solvency and those assessing risks for pricing in health and life insurance. Substantively, the risk assessment of the management system has a holistic approach *versus* the more selective focus on fundamental rights only of the impact assessment. The more global approach of the management system is more focused on the design, commercialisation and post-commercialisation of the AI-based tool, while impact assessment is focused on the concrete deployment or use of an AI within a company or public body.

Third. The lack of an ideal model for a risk management system in the field of AI classified as high risk. Available models accepted by the market and in comparative perspective. Throughout this paper we have analysed that the AIA gives a wide margin of discretion to the AI system provider to determine how to satisfy and set the minimum requirements and components of the AI risk management system, taking into consideration the current state of the art and scientific and technological developments. The objective, to have a management system, can be achieved in different ways: one can follow the components, requirements, and methodology of certain international standards such as ISO/IEC 42001:2023, ISO/IEC TR 24030:2021 and ISO/IEC TR 5469:2024; or the models of the technical standard NIST 800 218 PW.I.I and NIST 800 218RV.1.I.; or the OECD AI risk management model called *High-level AI risk-management interoperability framework*, among many others. Even if the above models and methodologies are accepted, they will have to be adapted to the context of each organisation, to the scope and nature of the AI lifecycle in question and to the data processing operations (database feeding the algorithm, extraction, training data, etc.). There is therefore no single or 'ideal' risk management model. Moreover, AI risk management systems must be coordinated with the other management models existing in the organisation (quality, environmental risks, information security, security schemes, etc.), to the extent that in some cases there may be an "integration" of the AI risk management system within other specific management models (in particular those of information security or compliance).

DATA AND DATA GOVERNANCE AND CONNECTIONS TO DATA PROTECTION PRINCIPLES IN ARTICLE 10 OF THE ARTIFICIAL INTELLIGENCE ACT

María Loza Corera

PhD in Law. Lead Advisor at Govertis part of Telefónica Tech.
Lecturer at the International University of La Rioja

I. Introduction

In today's world, nothing can be understood without data, not even the past. Data is an essential asset. In the context of the so-called digital economy, data play a role of paramount importance, to the point of talking about the data economy or data-driven economy[1]. In this context, Artificial Intelligence has even been mentioned as one of the most valuable intangible assets of any company as a driver of organisational value[2]. However, technology is not neutral[3], nor is the approach to risk regulation used[4], so design and data are absolutely relevant and the AI Act (hereafter AIA) is proof of this. The consequences of not having the right type of data, nor the required quality, could be disastrous, as they condition the results of the specific AI solution adopted from the design stage, and are therefore invalid and, more importantly, could affect the security and/or fundamental rights of individuals. The relationship between the data and the AI system is therefore directly proportional to the quality of the results obtained. However, not only will it be necessary to have adequate data sets and quality, but it is also essential to relate these data to the appropriate technology, specific internal procedures and for certain purposes determined by the organisation, not forgetting compliance with the different applicable regulatory frameworks, in other words, to establish a system of governance. At a time when we are already talking about the transition to the

[1] Loza Corera, M., De los microdatos a los datos masivos. Cuestiones legales, University of Valencia, 2017, p. 259.

[2] Witzel M. and Bhargava N., "AI-Related Risk The Merits of an ESG-Based Approach to Oversight", CIGI Papers No. 279, August 2023. https://www.cigionline.org/static/documents/no.279.pdf

[3] Floridi, L., "On Good and Evil, the Mistaken Idea That Technology is Ever Neutral, and the Importance of the Double-charge Thesis". Philosophy & Technology, September 2023, available SSRN: https://ssrn.com/abstract=4551487

[4] Kaminski M., "Regulating the risk of AI", 2022, Boston University Law Review, Vol. 103:1347, 2023, U of Colorado Law Legal Studies Research Paper No. 22-21, available SSRN: https://ssrn.com/abstract=4195066 p. 1351.

quantum economy[5], it is a *conditio sine qua non* to establish an adequate governance system to enable this transition.

The term Governance could well be one of those "*suitcase words*" that Marvin Minsky defined as words with multiple meanings. For this reason, it is necessary to clarify the different meanings of this term, although, as we will see, they are fully related, especially in the field of Artificial Intelligence. First, the concept of data governance will be addressed, taking into account the vital importance of data for an Artificial Intelligence system. Subsequently, the importance of data governance in the current European regulatory context and its meaning in this context will be analysed. Finally, it will analyse the concept of data governance in AIA, which is closer to the concept of *data equity*.

The AIA devotes an entire article (Article 10) in Chapter III, dedicated to high-risk AI systems, to *Data and data governance*, aware of the vital importance of data and data governance in an AI system. We can state without any doubt that this is one of the core articles of the Regulation, as not having adequate data sets will prevent the implementation of an AI system from the outset, not only because of the possible biases inherent in the underlying data, but also because AI also *learns* from data. The evolution, processing and final content of Article 10 will be studied in detail, including all the changes and modifications that have occurred since the European Commission's Proposal for a Regulation in April 2021, through the text proposed by the Council and the amendments approved by the European Parliament in June 2023, to its final version. It should be noted that the issues of accuracy, robustness and bias are not dealt with in this chapter, as they are specifically addressed in the chapter headed by Ana Aba Catoira. The above detailed study of the data governance obligations established by the Regulation will allow us to critically approach the final version contained therein.

Finally, we will briefly analyse the relationship between data governance and the principles of data protection, without prejudice to the more extensive general analysis of data protection in the chapter headed by Jesús Jiménez López.

II. Data governance

1. Concept of data governance

The concept of governance is not exclusive to data management, but

[5] World Economic Forum, "Quantum Economy Blueprint", January 2024, available at https://www3.weforum.org/docs/WEF_Quantum_Economy_Blueprint_2024.pdf.

rather has its origins in other areas, such as Information Technology Governance (IT Governance). However, given the increasing prominence that data has acquired in organisations, both public and private, the concept of data governance, proportionally to this prominence, has acquired its own substance and has become a real necessity.

Data is a core element in the digital economy, to the point of talking about the "data economy", so that properly managing this business asset is a necessary budget in order to be a *data-driven* company. Extracting value from data in order to make more conscious and effective decisions is a possibility that cannot be ignored in the current economic and technological context. This is where the concept of data governance takes on its full importance.

There is no unambiguous or normative definition for the concept of "Data Governance". Initially, data governance was understood to refer to the internal context of an organisation, only in relation to the control and management of its data, and has subsequently evolved into a broader and more elaborate concept. Thus, the *Data Governance Institute* defines it[6] as "the exercise of decision-making and authority in data-related matters", and more broadly, as "a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods".

For its part, the *Data Management Association (*DAMA) has created a reference framework for data management, *Data Management Body of Knowledge* (DMBOK[7]), in which data governance occupies an essential place within data management, making it clear that these are not overlapping concepts. That is why data governance or data governance is conceived as the "exercise of authority and control (planning, monitoring and enforcement) over the management of data assets".

The Spanish Data Protection Agency (AEPD) defines[8] data governance as "the strategy for the correct administration and management of data policy in the organisation". The AEPD stresses that the data protection policies to be adopted by the controller in compliance with Recital 78 and Article 24 of

[6] https://datagovernance.com/the-data-governance-basics/definitions-of-data-governance/

[7] The DMBOK focuses on eleven main themes: Data Governance; Data Architecture; Data Modelling and Design; Data Storage and Operations; Data Security; Data Integration and Interoperability; Documents and Content; Master and Reference Data; Data Warehousing and Business Intelligence; Metadata and Data Quality.

[8] AEPD, "Governance and Data Protection Policy", 2020 https://www.aepd.es/prensa-y-comunicacion/blog/gobernanza-y-politica-de-proteccion-de-datos

the General Data Protection Regulation[9] (GDPR) are an important part of the organisation's data policy.

It also indicates that, where personal data are processed, they should be added to the data governance objectives:

- Comply with data protection principles.

- To ensure that data subjects are able to exercise their rights.

- Ensure protection of personal data protection by design and by default, through risk management for rights and freedoms.

- Comply with the remaining legal obligations derived from data protection regulations.

Salvador Serna[10] highlights that, despite the multiple approaches to the concept of data governance, "there is a certain consensus in associating data governance with the ideas of: (1) valuing data as an asset of the organisation that must be managed (2) establishing responsibilities in decision-making (rights) and associated tasks (duties) and (3) establishing guidelines and standards to ensure the quality of data and its proper use". To these characteristics, we add a fourth, the need for strategic leadership from management for the establishment of a data governance system, not depending on an exclusive department or area of the company, so that, as a transversal system, it is coherent with the objectives and culture of the organisation and, of course, with the regulations in force.

It is therefore essential to have a data governance system in place, as it enables the comprehensive management of data throughout its life cycle, both in terms of quality, protection, security and maintenance, as well as regulatory compliance. In addition to obtaining the maximum value from the data to help in making more efficient decisions, proper data governance minimises risks, saves costs by centralising information management, eliminates silos, improves data quality and processes thanks to the monitoring and continuous improvement system and, very importantly, establishes the conditions to allow the scalability of different AI solutions that can be adopted. Therefore, we move from the concept of data governance to AI governance, but the former being a necessary presupposition to be able to talk about the latter.

[9] REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of individuals with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation).

[10] Salvador Serna, M., (2021), Inteligencia artificial y gobernanza de datos en la Administración Pública: sentando las bases para su integración a nivel corporativo, in *Repensando la administración pública: administración digital e innovación pública,* (pp. 126-148), INAP, 2021.

Industry is well aware of the imperative need for an AI governance system, not only to comply with regulations, but to drive business value.[11]

2. European context

The above concept of data governance, which can be referred to as 'micro', must necessarily be put in relation to the current EU political and regulatory context, in particular, to the data governance mechanisms or regulatory requirements at the 'macro' level necessary to enable the single market for data.

In 2018 the European Commission launched its *Artificial Intelligence Strategy*[12] where it laid the foundations to ensure that the potential of AI serves human progress by enhancing the Union's technological and industrial capacity, by preparing for the socio-economic transformations that AI will bring about, and by establishing an appropriate ethical and legal framework, based on the Union's values and in line with the EU Charter of Fundamental Rights. In this way forward, a clear and essential objective is to increase the volume of data available and to facilitate access to it. Thus, the European Commission, aware of the value of data for both the economy and society and, without renouncing the protection of personal data, has promoted the *EU Data Strategy*[13], in the framework of the policy priorities set for the period 2019-2024 (*A Europe fit for the digital age*)[14] and of the *Digital Compass 2030: Europe's approach for the Digital Decade.*[15]

In the *European Data Strategy* the Commission states that "The aim is to create a single European data space, a true single data market, open to data from all over the world, where personal and non-personal data, including sensitive business data, is secure and where businesses also have easy access to an almost infinite amount of high quality industrial data, in a way that drives growth and creates value, while minimising the human environmental and carbon footprint".

To achieve such a single European data space that ensures Europe's glob-

[11] IBM, The urgency of AI governance, 2023. https://www.ibm.com/downloads/cas/MV9EXNV8

[12] COM(2018) 237 final, *Artificial Intelligence for Europe.*

[13] COM(2020) 66 final, *A European Data Strategy,* European Commission https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:52020DC0066

[14] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit--digital-age_es

[15] COM(2021) 118 final https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:52021DC0118

al competitiveness[16] and data sovereignty[17], as stated in the European Data Strategy, EU legislation must be effectively implemented so that all data-based products and services comply with the rules of the single data market. Alongside appropriate legislation, 'clear and reliable' governance mechanisms to enable access to and use of data must be adopted to ensure that the objectives of the European data space are met.

The European regulatory framework designed to enable the realisation of the European Data Strategy consists, among others, of the Data Governance Regulation 2022/868[18] and Regulation 2023/2854 on harmonised rules for fair access to and use of data (Data Regulation)[19], not forgetting the Regulation on a framework for the free flow of non-personal data in the European Union[20], consistent with the meaning given to the concept of 'data' by the above-mentioned Regulations, whose meaning is much broader than the concept of 'personal data'.

It should be emphasised that the European single data market is not unaware that international data flows are indispensable in today's markets and competitive environments, and therefore has an open approach, but without renouncing European protection and values.

We see, therefore, how we have progressively evolved from a regulation focused on the protection of personal data and the rights and freedoms of individuals, to a strategy focused on data (not necessarily personal) as a business asset, the centre of the data economy, which needs regulations that guarantee its availability, sharing and secure reuse, but always preserving European values. This is why, in order to guarantee the single market for data (governance at the "macro" level), it is essential for organisations to have solid data governance at the internal level (micro level), which will also make it possible to move towards the governance of Artificial Intelligence.

[16] COM(2020) 66 final "However, the sources of competitiveness for the coming decades in the data economy are determined now. This is why the EU must act now".

[17] The functioning of the European data space will depend on the EU's ability to invest in the next generation of technologies and infrastructures, as well as in digital skills such as data literacy. This, in turn, will increase Europe's technological sovereignty in terms of key enabling technologies and related infrastructures for the data economy.

[18] Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Regulation).

[19] Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules for fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Regulation).

[20] https://eur-lex.europa.eu/legal-content/ES/TXT/HTML/?uri=CELEX:32018R1807&qid=1696786250350

3. Concept of governance in the field of Artificial Intelligence

The AIA does not provide a definition of governance applied to the field of AI. Nor is there a definition in ISO *Information technology-Artificial Intelligence - Artificial Intelligence concepts and terminology ISO/IEC 22989*. The IAPP[21] defines "*AI Governance*" as "*A system of policies, practices and processes organisations implement to manage and oversee their use of AI technology and associated risks to ensure the AI aligns with an organisation's objectives, is developed and used responsibly and ethically, and complies with applicable legal requirements*". Similarly, the industry defines it as "*AI governance is a system of rules, practices, processes and tools that help an organisation use AI in alignment with its values and strategies, address compliance requirements and drive trustworthy* performance"[22]. It is argued that AI governance is likely to be as important as the specific governance of the components of the algorithm itself.[23]

However, regardless of whether there is a normative definition or not, it is unquestionable that the concept of governance takes on its full importance in the field of AI to the point of transcending the concept of data governance to speak of AI governance. Any organisation must establish the necessary procedures to ensure compliance with applicable regulations, the necessary security measures and respect for fundamental rights and freedoms, as well as to guarantee the proactive responsibility of the organisation and its governing bodies in the use of the different AI solutions it decides to implement. In fact, if we had to sum up AIA in one word, it would be "Governance".

It should not be misunderstood that these obligations only fall on the entities that develop AI systems, but that those that design or deploy them (deployers or those responsible for the deployment) also have responsibilities, so that, although at different levels, it is necessary for all organisations to establish AI governance mechanisms.

There are different AI governance systems or frameworks. In the field of *soft law*, the *Artificial Intelligence Risk Management Framework*[24] (AI RMF) of the *National Institute of Standards and Technology* (NIST) in the US and the *Governance Guidelines for the Implementation of AI Principles*[25] in Japan stand out, al-

[21] IAPP, Key Terms for AI Governance, June 2023. https://iapp.org/resources/article/key-terms-for-ai-governance/

[22] Op. cit. IBM, The urgency of AI governance, 2023.

[23] In fact, in 2022, AI governance was the ninth most important strategic priority for privacy functions. In 2023, it is the second most important strategic priority, IAPP-EY Professionalizing Organizational AI Governance Report, p. 9, 2023.

[24] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[25] AI Governance in Japan, REPORT FROM THE EXPERT GROUP ON HOW AI

though to date there is no binding regulatory framework in this area, whilst there is already a glimpse of its forthcoming approval in both countries. By contrast, in Europe there was no specific framework for AI governance until the adoption of the European Regulation.

Regardless of the different approaches to AI Governance, the very concept of Governance is fully in line with Novelli, Taddeo and Floridi's[26] assertion that proactive accountability is a cornerstone of AI governance.

The AIA refers specifically to data governance. Thus, both Recital 67 and Article 10 refer to "good governance and data management practices", which we will analyse below. Therefore, leaving aside Chapter VII dedicated to institutional governance at both the European (European Artificial Intelligence Committee) and national (national competent authorities) levels, the Regulation refers to the concept of data governance, thus at the 'micro' level. This does not mean at all that the AI governance established by the European Regulation is exhausted in this article rather dedicated to data governance, but it must be put in relation with the other obligations established for high-risk AI systems, which require the implementation of other procedures, such as conformity assessment procedures, declaration of conformity and CE marking, quality management systems that include compulsory change management procedures, techniques, procedures and systematic actions to be used for design, design control and design verification and quality control, data management systems and procedures, risk management system, post-market surveillance, serious incident reporting procedure, procedures for recording all documentation and the establishment of an accountability framework. All these obligations make up what we mean by AI governance under the European Regulation.

Finally, there is a broader perspective on AI governance, directed at regulators, in that some authors consider that the regulation being proposed in Europe, Canada and elsewhere is not sufficient to prevent other risks that may occur in the longer term. Thus, KOLT[27] argues that regulatory proposals

PRINCIPLES SHOULD BE IMPLEMENTED, 2021, available at https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20210709_8.pdf.

[26] NOVELLI C., TADDEO M., FLORIDI L., Accountability in artifcial intelligence: what it is and how it Works, AI & Soc (2023) https://doi.org/10.1007/s00146-023-01635-y

[27] Kolt, N., Algorithmic Black Swans (October, 2023). Washington University Law Review, Vol. 101, Forthcoming, available SSRN: https://ssrn.com/abstract=4370566 p. 42. These principles are: Principle 1: AI governance should seek to anticipate and mitigate large-scale harms from AI systems; Principle 2: AI governance should adopt a portfolio approach composed of diverse and uncorrelated regulatory strategies; Principle 3: AI governance should be highly scalable; Principle 4: AI governance should continuously explore and evaluate new

to regulate AI "focus primarily on the immediate risks of AI, rather than on broader, longer-term risks" and therefore "offers a roadmap for "algorithmic preparedness": a set of five forward-looking principles to guide the development of regulations that address the prospect of algorithmic black swans and mitigate the harms they pose to society".

III. Development, processing and final content of Article 10

The AIA dedicates Article 10, within Section 2 of Chapter III dedicated to High Risk AI Systems, to *Data and data governance*, aware of the vital importance that data and data governance have within an AI system. To approach the analysis of its content, we will first analyse the roles involved, and then delve into the obligations associated with each of them.

1. Roles involved

It should be emphasised that the obligations set out in Article 10 relating to data and data governance are set out without mentioning the specific data subject, as they are configured as requirements of the high-risk system itself.

Therefore, in order to establish which parties are obliged to implement good data governance and management practices, we must first look at the datasets on which these obligations fall, to see to which part of the value chain they correspond. Thus, Article 10 distinguishes between datasets used for training, validation, and testing of high-risk AI systems, as distinct from those that do not use techniques involving model training. [28]If we look at the various figures that make up the value chain, already analysed throughout this work, leaving aside those roles that do not directly influence the development of the AI system, such as the distributor[29] or the importer[30], the provider[31] stands out. The provider means the entity that *develops or for which* an AI system or a general-purpose AI model *is developed* and brings it to the market or puts it into service under its own name or trademark, whether for payment or free of charge, and should therefore have data sets used for training, validation and testing of the system. However, the provider, by definition, can either de-

regulatory strategies; Principle 5: Cost-benefit analysis of AI governance interventions should take more account of worst-case outcomes.

[28] Article 10.6.

[29] Article 3. 7).

[30] Article 3. 6).

[31] Article 3(3).

velop the AI system directly or contract third parties to carry out such development, in which case, the corresponding obligations must be contractually regulated. In this regard, the European Parliament recalled in its Recitals[32] that algorithm *developers* are particularly relevant as they may have used underlying (historical) data which may not meet the desirable quality requirements due to biases, or may have generated this data in real environments and therefore be biased by default. Finally, the explicit reference to developers has been omitted from the Recital, while maintaining the importance of the quality of the underlying data.

In relation to the figure of the *provider*, it should be borne in mind that, in terms of responsibilities along the AI value chain, the Regulation in certain cases considers[33] "provider" of a high-risk AI system to be any distributor, importer, deployer or other third party and, therefore, subject to the obligations set out in Article 16 for providers and deployers of high-risk AI systems and other parties. In this regard, it should be noted that the Parliament proposed[34] to amend the title of Article 16 to include not only providers, but also deployers and other parties, but this amendment was not finally accepted. However, despite the obvious coherence of the amendment proposed by the Parliament, this does not affect the substance, as Article 25.1 expressly provides for the liability of these figures. Therefore, any distributor, importer, deployer, or other third party who (i) places its name or trade mark on a high-risk AI system already placed on the market or put into service (ii) makes a substantial modification or (iii) makes a modification in such a way that the AI system becomes a high-risk AI system, will be subject to the obligations set out in Article 16 and thus to compliance with all the requirements for high-risk systems, including those relating to data governance. The final version[35] has included the definition of "downstream provider", defined as a provider of an AI system, including a general purpose AI system, which integrates an AI model, regardless of whether the model is provided by themselves and vertically integrated or provided by another entity based on contractual relationships.

For its part, the *deployer*[36] is the entity that uses an AI system under its own

[32] Amendment 78 on Recital 44 (now Recital 67).

[33] Article 25.1.

[34] Amendment 331, Article 16, title: "Obligations of providers and deployers of high-risk AI systems and other parties".

[35] Article 3.68.

[36] Article 3.4. It is worth highlighting the relevant change introduced by the European Parliament (through Amendment 172 which modifies the definition of user in Article 3.4) in coherence with Recital 59) which dispenses with the term "user" to call it "deployer", which

authority, provided that the "domestic exception" does not apply, meaning this that its use is part of a personal activity of a non-professional nature. Although not expressly mentioned, we understand that the deployer will be liable whenever he retrains the system given by the provider. The issue of retraining will be discussed in more detail below.

Article 10 only expressly mentions the provider in relation to the possibility to exceptionally process special categories of data to the extent strictly necessary to ensure the detection and correction of negative bias. The Parliament added[37] a second express reference to the provider, in setting out its possible exemption from liability for breach of any of the obligations laid down in Article 10, transferring such liability to the deployer, in case the provider does not have access to the data, because they are held exclusively by the deployer and this has been laid down in a contract. This paragraph has not been included in the final version, but it is questionable what sense it would make for a deployer to have exclusive access to the data of a system introduced to the market by a provider, but without the deployer using it under their own authority, as in that case they would already have responsibility for it.

In any case, of particular interest is the mention[38] made by the Parliament in relation to the possibility of outsourcing the requirements related to data governance "by using third parties offering certified compliance services, including verification of data governance, data set integrity and data training, validation and testing practices", which has been accepted in the final text. Therefore, we believe that a new figure ("data verifiers" or "certified data service providers") will enter the value chain, precisely in charge of supplying providers or deployers with datasets for the development of AI systems, which comply with the requirements established by the AIA.

2. Obligations

Article 10 on data and data governance is of paramount importance [39] as compliance with the obligations set out therein is the basis for high quality

we understand to be more clarifying as it rules out confusion with the end user of the system, a natural person.

[37] Amendment 291 introducing a new paragraph 6a.

[38] Amendment 78 modifying Recital 44 *in fine* (now Recital 67).

[39] Recital 67 states (unofficial translation) "High quality data and access to high quality data play an essential role in providing structure and ensuring the functioning of many AI systems, in particular where techniques involving model training are employed, with a view to ensuring that the high-risk AI system operates as intended and safely and does not become a source of any discrimination prohibited by Union law (…)".

data and thus for the proper functioning of AI systems, especially high-risk ones.

Thus, high-risk AI systems that make use of techniques involving the training of models with data, should be developed from datasets that meet the *quality criteria* specified in paragraphs 2 to 5. In contrast to the requirement to use training, validation and test datasets that meet the above quality criteria, it should be noted that the European Parliament proposed a modulation[40] that these quality criteria should be met "to *the extent that this is technically feasible* in accordance with the market segment or scope of application concerned". The Parliament also pointed out that these criteria should be met for techniques that do not require labelled input data, such as unsupervised learning and reinforcement learning. Neither of these two proposals of the Parliament was finally accepted, so that the final version has eliminated any kind of modulation of responsibility. We will now look at the quality criteria set out in each of the paragraphs.

The second paragraph specifies the *good* data governance and management *practices* to which training, validation, and test datasets used for training models of high-risk AI systems should be subjected. We can classify these practices around different actions:

- (a) relevant design decisions;

- its recompilation: (b) data collection processes and the origin of data, and in the case of personal data, the original purpose of the data collection;

- data preparation: (c) relevant data-preparation processing operations, such as annotation, labelling, cleaning, updating, enrichment and aggregation;

- (d) the formulation of assumptions, in particular with respect to the information that the data are supposed to measure and represent;

- to the preliminary study of the datasets: (e) an assessment of the availability, quantity and suitability of the data sets that are needed);

- the quality of the data:

- (f) examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations;

- (g) appropriate measures to detect, prevent and mitigate possible biases identified according to point (f); and

- (h) the identification of relevant data gaps or shortcomings that prevent compliance with this Regulation, and how those gaps and shortcomings can be addressed.

[40] Amendment 278 modifying Article 10(1).

Firstly, it should be noted that, unlike the Commission and Council proposals which spoke of "good governance and data management *practices*", the Parliament[41] proposed to replace this term by "*appropriate governance* to the context of use as well as the intended purpose of the AI system" which implied the adoption of a series of *measures*. The final version does not take up the Parliament's proposal and reverts to "appropriate governance and data management *practices* fit for the intended purpose of the AI system".

As regards the specific practices, most of the amendments introduced by the Parliament have finally been accepted. Thus, the Parliament included[42] a new practice concerning transparency on the original purpose of data collection, which the final text[43] further specifies by distinguishing between processes for the collection of non-personal data, in which case the origin of the data must be indicated, and personal data, for which the original purpose of collection must be indicated.

In the practice concerning the preparation operations[44] of the data, the Parliament added the update[45] of the data.

Regarding the prior study of the datasets[46], the Parliament[47] removed the requirement for prior assessment of the availability, quantity, and adequacy of the necessary datasets, which in our view does not make much sense, as, although such an assessment should obviously be prior, it reinforced its *ad hoc* character.

But it is in the measures relating to data quality that the Parliament made the most significant changes. Thus, as regards the examination of possible bias[48], the Council added the precision that they could "affect the health and safety of natural persons or give rise to discrimination prohibited by EU law". For its part, the Parliament added[49] that they could "adversely affect fundamental rights". In relation to that they may give rise to discrimination prohibited by EU law, the Parliament added "in particular where the outgoing data influence the incoming data in future operations ("feedback loop"), a clarification that has not been included in the articles", but has been included

[41] Amendment 279 modifying Article 10(2).

[42] Amendment 280 including a new paragraph (aa).

[43] Article 10.2(b).

[44] Article 10(2)(c).

[45] Amendment 282.

[46] Article 10.2(e).

[47] Amendment 284.

[48] Article 10(2)(f).

[49] Amendment 285.

in Recital 67. The Parliament also introduced a new practice[50], consisting of carrying out "appropriate measures to detect, prevent and mitigate potential biases", which goes beyond *ex ante* examination of particular datasets and requires measures to be put in place to detect, prevent, and mitigate potential biases that may be detected or become apparent at a later stage.

In relation to the practices concerning the identification of possible data gaps or deficiencies and how to remedy them[51], the Parliament introduced[52], and this is reflected in the final version, the qualification that such gaps or deficiencies shall be those "relevant to prevent compliance with this Regulation", thus seeming to narrow the objective scope of these remedyable gaps or deficiencies.

The third paragraph of specifies a series of *obligations* ***that*** began as a result, but which have finally been modulated. Thus, it states that the data sets used for training, validation and testing "shall be relevant, sufficiently representative and, as far as possible, error-free and complete for the intended purpose".

In relation to this first obligation, the Council introduced a first modulation by including "as far as possible" before the imperative ("shall be free of errors and complete"). Subsequently, the Parliament[53] significantly modified the wording and added the adverb "sufficiently" representative to the obligation for the data to be representative.

Secondly, the obligation of result to be free of errors and complete is transformed by the Parliament into "duly assessed for errors and as complete as possible in view of the intended purpose"[54]. The final text similarly states "to the best extent possible free of errors and complete in view of the intended purpose of the system". This development can also be seen in a correlative manner in Recital 67.

Finally, an obligation is added that the datasets shall have appropriate statistical properties, in relation to the persons or groups of persons for whom the high-risk AI system is intended to be used. The datasets *may meet* these characteristics individually for each data item or for a combination of data items. The Parliament corrects that the datasets *shall meet* these characteristics,

[50] Amendment 286 which introduced a new paragraph (f a), now paragraph (g).

[51] Article 10(2)(h).

[52] Amendment 287.

[53] Amendment 288.

[54] Consistent with Recital 44 which states that "(…) they should be sufficiently relevant and representative, adequately checked for errors and as complete as possible in view of the intended purpose of the system (…)".

not individually for each data item, but for each dataset or for a combination of datasets, as the final text reads.

The fourth paragraph states that the data "ss shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, contextual, behavioural or functional setting within which the high-risk AI system is intended to be used". The Parliament proposed to add[55] that reasonably foreseeable misuses of the AI system should also be taken into account, which will not be taken up by the final version. On the other hand, this obligation should be connected to the presumption set out in Article 42 whereby the requirements of the fourth paragraph shall be presumed to be met provided that high-risk AI systems have been trained and tested with data reflecting the specific geographical, behavioural, contextual, and functional environment in which they are intended to be used.

The fifth paragraph establishes the possibility for providers to process special categories of data "to the extent that it is strictly necessary for the purpose of ensuring bias detection and correction". It should be noted that the Parliament called such *negative* biases and defined them[56] as those that "create(s) a direct or indirect discriminatory effect against a natural person", but in the end, the concept of "negative bias" was not taken up in the final version. It provides for the possibility of establishing an adequate legitimation basis[57] in order to be able to process special categories of data which, in application of data protection law, shall not exempt from the obligation to adopt appropriate safeguards

the rights and freedoms of natural persons. Recital 70 expressly speaks of 'a matter of substantial public interest' and rescinds the express reference to Article 9.2(g) of Regulation (EU) 2016/679 and Article 10.2(g) of Regulation (EU) 2018/1725, which the Council had first introduced. On this point, for data processing to be covered by Article 9.2(g) of the GDPR (processing is necessary for reasons of essential public interest), it should be recalled that this must be provided for in a rule of national or European law, which also specifies the essential public interest justifying the processing of such data, in which circumstances the right to data protection may be limited, precise rules and appropriate safeguards at both technical and organisational level to

[55] Amendment 289.

[56] Amendment 78 in Recital 44 *in fine*.

[57] Amendment 160 introducing a new Article 2.5a: "This Regulation shall not affect Regulation (EU) 2016/679 (…), without prejudice to the mechanisms provided for in Article 10(5) (…)" which is finally included in the final text.

protect the interests and fundamental rights of the data subject. Here, the Parliament, instead of listing by way of example a number of measures, introduced[58] a catalogue of necessary conditions that must apply for processing to take place, including that the processing of synthetic or anonymised data does not effectively achieve the detection and correction of bias; that the data to be used are pseudonymised or subject to technical limitations on the re-use of personal data and to the most advanced security and privacy-preserving measures; or that they are deleted once the bias has been corrected or when the personal data reach the end of their retention period, which have been set out in the final text.

The European Parliament underlines the exceptionality of the fact that providers of such systems may process special categories of data by introducing the adverb 'exceptionally'. In this regard, the Parliament introduced a requirement that providers making use of this provision should produce documentation explaining why the processing of special categories of personal data is necessary to detect and correct bias. In the final version this obligation does not expressly mention providers and merely states that records of processing activities in accordance with Regulation (EU) 2016/679, Directive (EU) 2016/680 and Regulation (EU) 2018/1725 should include such justification.

Having seen the quality criteria established for training, validation, and test datasets for the development of high-risk AI systems using techniques involving model training with data, the sixth paragraph establishes that these quality criteria, with respect to the development of high-risk AI systems that do not employ techniques involving model training, shall only apply to test datasets. Interestingly, while the Commission stated for these systems that they should ensure compliance with the good data governance and management practices set out in the second paragraph, the Council, the Parliament and the final version extend this obligation to all quality criteria (paragraphs 2 to 5) but limit such compliance only to test datasets.

IV. Critical approach

In the light of the development of the normative proposal for Article 10, this section will critically assess the final content of Article 10.

Firstly, in relation to roles, a general criticism is the absence of a definition of end-user or 'affected' by the AI system, especially considering that

[58] Amendment 290.

the Parliament proposed to introduce a definition[59] of 'affected person' and that these fall within the scope of the AIA[60]. In relation to the AI value chain and the roles involved in it, following the important reference[61] introduced by the Parliament in relation to the possibility of outsourcing the requirements related to data governance, as discussed in the section on roles, we understand that all the circumstances are in place for the emergence in the value chain of new figures that exclusively provide "verified" data and certify that the data comply with the established governance requirements, as well as their integrity and training ("data verifiers" or "certified data service providers"). Regulation and possible transfer of responsibility will therefore be key. However, is questionable whether this model of "verified" data provision can deliver the individualised compliance with such governance requirements that the Regulation aspires to, since, firstly, governance must be tailored to the context of use as well as to the intended purpose of the AI system[62] and, secondly, the sets of data sets that will be used for the purpose of the AI system will need to be defined, secondly, datasets should take into account, to the extent required by the intended purpose, the characteristics or elements specific to the geographical, behavioural, contextual, or functional environment in which the high-risk AI system is intended to be used.[63] In this way, as PEGUERA POCH[64] warns, the value chain could acquire "different configurations to those considered by the legislator depending on the evolution of the business models that end up being consolidated".

Moreover, as recommended by the European Data Protection Supervisor (EDPS)[65], it should be specified that AI operators who retrain pre-trained AI systems are included in the concept of providers, as AI systems may be

[59] person concerned: any natural person or group of persons who are exposed to an AI system or otherwise affected by an AI system". The reasons which may have led to the non-acceptance of this amendment are not understood, especially when it does introduce the reference to the mechanisms of guarantee or protection in the event of infringement of the Regulation, which, on the other hand, are not reserved to the person affected by the AI system, but to any person who considers that there has been an infringement of the Regulation; and, finally, because it does provide other definitions which do not seem so relevant, such as the "subject" who participates in tests under real conditions or the "informed consent" of this person.

[60] Article 2.1(g).

[61] Amendment 78 modifying Recital 44 *in fine*.

[62] Article 10, second paragraph.

[63] Article 10, fourth paragraph.

[64] Peguera Poch, M. "La propuesta de reglamento de AI: una intervención legislativa insoslayable en un contexto de incertidumbre", in Peguera Poch (coords.) *Perspectivas regulatorias de la Inteligencia Artificial en la Unión Europea*, Madrid: Reus, 2023.

[65] EDPS, *Opinion 44/2023 on the Proposal for Artificial Intelligence Act in the light of legisla-*

trained more than once during their lifecycle or may apply continuous learning techniques. Retraining may be due, says the EDPS, either to the lack of large data sets for training, or because they are retrained in order to be used for a similar task in a different domain (transfer learning). The AIA also does not clarify whether retraining or continuous learning activities are considered as part of the 'development' of the AI system, as in that case they would clearly be considered as providers. The EDPS states that this point is particularly relevant in relation to foundational models and the generalised possibility of retraining them. The Regulation does not include a definition of the operations that are included in the 'development' of an AI system, and in the definition of provider, although it includes a reference to the development or marketing under its name or brand of a general purpose AI model, there is no mention of retraining. However, only one Recital[66] expressly mentions retraining as a process that can be incorporated by the provider into the AI system. Therefore, a systematic and teleological interpretation would lead us to consider the provider as the one who introduces a retrained system on the market, although the clarification made by the EDPS would have been appropriate.

Secondly, Article 10 refers to the requirements to be met by training, validation and test datasets to be used for the development of high-risk AI systems using techniques that involve training models with data. It has been highlighted by certain authors[67] that it ignores other stages of *machine learning that* should also be subject to data quality criteria and data governance practices and also with respect to data licences that allow access to data.

The first paragraph establishes a sort of obligation of result, stating that high-risk AI systems that make use of techniques involving the training of models with data "shall be developed on the basis of training, validation and testing data sets that meet the quality criteria referred to in paragraphs 2 to 5 (…)". The Parliament proposed to introduce a liability modulation, or rather the removal of such an obligation of result in respect of all governance ob-

tive developments, p. 8. https://edps.europa.eu/system/files/2023-10/2023-0137_d3269_opinion_en.pdf

[66] Recital 88: 'Within the AI value chain, multiple parties often provide AI systems, tools and services, but also components or processes that are incorporated by the provider into the AI system for various purposes, including model training, model *retraining*, model testing and evaluation, integration into software or other aspects of model development (…)'.

[67] Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H., & Steinrötter, B. "The european Commission's proposal for an Artificial Intelligence Act-A critical assessment by members of the robotics and AI law society (RIALS)", 2021, J, 4(4), p. 595. doi: https://doi.org/10.3390/j4040043

ligations, by making compliance with such requirements "technically feasible in accordance with the relevant market segment or scope ". This modulation was a significant modification, but in practice it might not be so, since it was based on purely technical criteria, and the justification for the impossibility of complying with some of the required quality criteria would have to demonstrate precisely the technical impossibility in each specific case. What is relevant is that the final version has eliminated any kind of modulation of liability, regardless of the specific segment or scope of application or technical impossibility, which reinforces the importance of complying with the quality criteria in any case.

The second paragraph introduces the governance and management practices to be complied with by the data sets for training, validation, and testing of high-risk systems, which involve a whole data management system. These data governance practices must necessarily be connected to the quality management system and, in particular, to the risk management system, although this is not expressly stated, which would have been desirable, as it would underline the importance of compliance with Article 10, which, as we have said, is essential. The quality management system does mention[68] "data management systems and procedures including data acquisition, collection, analysis, labelling, storage, filtering, searching, aggregation, preservation and any other data-related operations carried out before the introduction to the market or commissioning of high-risk AI systems", but we believe it would have been desirable to make express reference to the complete data governance system established in Article 10, in the same way as the express reference to the risk management system is included. In relation to the risk management system established in Article 9, it is stated that "The risks referred to in this Article shall concern only those which may be reasonably mitigated or eliminated through the development or design of the high-risk AI system, or the provision of adequate technical information", which seems to exclude risks arising from non-compliance with data quality criteria. However, the same article specifies that the known and foreseeable risks to health, safety, or fundamental rights that the high-risk AI system may entail should be identified and analysed, which implies that risks arising from non-compliance with data quality criteria and governance practices cannot be ignored. In any case, the data governance system set up by the AIA has sufficient substance of its own that it transcends the risk management system, but this does not imply that it is unknown to the latter.

The importance of the data governance system is evidenced by the fact

[68] Article 17.1(f).

that it forms part of the technical documentation (Annex IV) to be retained by the provider for ten years, although it does not explicitly mention Article 10, but refers, in relation to the data, to a general description of the training data sets used and information about their provenance, scope and main characteristics; the way in which the data were obtained and selected; the labelling procedures (e.g., for supervised learning) and data cleaning methodologies (e.g., anomaly detection); and the validation and test procedures used, including information about the validation and test data used, and its main characteristics. It would have been desirable to include an explicit reference to Article 10 data governance procedures in order to provide sufficient traceability for potential liability claims.

In the Commission's proposal, the third paragraph established an obligation of result that the data sets to be used for training, validation, and testing "shall be relevant, representative, error-free and complete". Industry or even some governments, such as the Norwegian government[69] and some authors, were reluctant to draft it as an "absolute requirement", as it is an impossible task that data can always be free of errors and such a level of perfection is "technically unfeasible" and could hamper innovation[70]. Other authors[71] have highlighted the existence of conditionalities to the fulfilment of these apparently strict obligations, which in fact lower the level of requirements. Thus, the successive versions have introduced formulas that have lowered the level of requirements for obtaining these results, so that the final version establishes that the data must be relevant, sufficiently representative *and, as far as possible*, free of errors and complete, taking into account the intended purpose. It would have been advisable to also introduce, together with the purpose, the reference to reasonably foreseeable misuses[72], for the sake of consistency as these are taken into account in the risk assessment of Article 9.[73]

This apparent modulation of responsibility, we believe, should be connected to the concept of proactive responsibility, so it must be possible to demonstrate relevance, sufficient representativeness, analysis of possible errors, and data completeness, although it is true that due to the very nature

[69] https://www.regjeringen.no/contentassets/939c260c81234eae96b6a1a0fd32b6de/norwegian-position-paper-on-the-ecs-proposal-for-a-regulation-of-ai.pdf

[70] *Cit*. Ebers, M. *et alia*.

[71] Veale M. and Borgesius F., "Demystifying the Draft EU Artificial Intelligence Act", Computer Law Review International, 2021, 22(4), pp. 97-112, para. 41. DOI https://doi.org/10.48550/arXiv.2107.03721

[72] Article 3. 13).

[73] Article 9.2(b).

of AI, it may be problematic to evaluate the responsibility for the results obtained.[74]

On the other hand, it is somewhat paradoxical to speak of "quality criteria" when no criteria for measuring the quality of the datasets are specified, referring only to the desirable outcome[75]. In other words, the AIA leaves such specification to the field of standardisation, which is in a way understandable as it deals with mostly technical aspects, but at the same time leaves the standard somewhat empty of substantive content[76]. Thus, it is stated[77] that "standardisation should play a key role to provide technical solutions to providers to ensure compliance with this Regulation(…)". It should be noted at this point that the amendments[78] made by the Parliament imply an active role for the Commission and not a mere "outsourcing" of the issue to standardisation bodies. Thus, the Commission, taking into account the importance of standards in ensuring compliance with the requirements of the Regulation and the competitiveness of enterprises, provides that in the development of standards there should be a balanced representation of interests by encouraging the participation of all relevant stakeholders. In order to facilitate regulatory compliance, the Commission should, no later than two months after the adoption of the AIA, issue the first requests for standardisation to the European standardisation organisations.[79]

At this point, it should be noted that the use of private bodies for the elaboration of standards is criticised by certain authors[80], especially when such apparently "technical" standards have an impact on fundamental values or rights. This is evident when the AIA[81] states that the Commission shall be empowered to adopt common specifications when the relevant harmonised

[74] Op. cit. Novelli C., Taddeo M., Floridi L., Accountability in artifcial intelligence.

[75] *Cit*. Ebers, M. *et alia* mention predictive accuracy, robustness and the unbiasedness of trained machine learning models as possible criteria.

[76] In the Proposed Standard Contractual Clauses for the procurement of Artificial Intelligence by public bodies, September 2023 version, Article 3 (characteristics of datasets) is exactly the same for high-risk AI systems as for all other systems. Available at https://public-buyers-community.ec.europa.eu/communities/procurement-ai/resources/eu-model-contractual-ai-clauses-pilot-procurements-ai

[77] Recital 121.

[78] Amendments 103 to 107 concerning Recital 61 (now Recital 121).

[79] CEN (European Committee for Standardisation), CENELEC (European Committee for Electrotechnical Standardisation) https://www.cencenelec.eu/

[80] Veale M., and Borgesius F., "Demystifying the Draft EU Artificial Intelligence Act- Analysing the good, the bad, and the unclear elements of the proposed approach", *Computer Law Review International*, vol. 22, no. 4, p. 105.

[81] Article 41(1)(a)(iii).

standards do not sufficiently address fundamental rights issues. It should be recalled that high-risk AI systems or general purpose AI models which are in conformity with harmonised standards to be adopted will be *presumed*[82] to comply with the requirements[83] set for high-risk AI systems, and therefore a procedure based on internal control (Annex VI), which does not foresee the involvement of a Notified Body, will suffice to obtain conformity assessment (Annex VI). Therefore, only where harmonised standards or common specifications do not exist or have not been implemented, a conformity assessment procedure involving a Notified Body (Annex VII) will be followed. It is the providers of these systems, before placing them on the market or putting them into service, who shall ensure that they have been subject to the appropriate conformity assessment procedure[84] and, if positive, shall draw up the EU declaration of conformity[85] and affix the CE marking[86]. It goes without saying that the "self-assessment" of conformity ultimately entails fewer guarantees precisely with regard to the verification of compliance with the requirements, let us remember, for high-risk AI systems, and it would therefore be desirable that the prior conformity assessment procedure for high-risk AI systems should always be carried out by a third party other than the provider. This point has also been called for by both the European Data Protection Board (EDPB) and the European Data Protection Supervisor (EDPS), which state [87] that, although the GDPR does not provide for an obligation to carry out a third party conformity assessment for high-risk data processing, the risks in the field of AI are not yet fully known. This is why they advocate introducing *ex-ante* third party conformity assessment in general, and not only for certain high-risk systems, as this would 'further enhance legal certainty and confidence in all high-risk AI systems'. The EDPS subsequently reaffirms[88] and adds that, taking into account the sectoral legislation applicable to the activity in the context of which the AI system will be used, the third party assessment of the high risk AI system, in order to ensure the reliability of the

[82] Article 40.1.

[83] Chapter IV.

[84] Article 43.

[85] Article 47.

[86] Article 48.

[87] EDPB-EDPS *Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act),* 18 June 2021, paragraph 37. Available at https://edps.europa.eu/system/files/2021-06/2021-06-18-edpb-edps_joint_opinion_ai_regulation_en.pdf.

[88] *Cit.* EDPS, *Opinion 44/2023,* para. 28.

AI, will require the involvement of the supervisory authority with specific expertise in the field.

Therefore, considers that it cannot be left to the provider's discretion whether or not to submit to third party verification as has been maintained both by the Parliament[89], and by the final text. On the other hand, we also fail to understand why the reference to the existence or not of harmonised standards or common specifications is only taken into account for one type of high-risk AI systems (specifically those related to biometric identification and categorisation of natural persons) and is not taken into account in a generalised way for all of them.

In relation to the requirement for data to be complete and, as far as possible, error-free, it has become clear that the use of techniques such as differential privacy implies the introduction of noise to avoid inadvertent disclosure of sensitive data. For this reason, some authors[90] advocate that Article 10 should allow the use of these privacy enhancing techniques (*PETs*) in the data governance practices of high-risk systems. In this regard, it should be noted that the Council introduced in Recital 44 (now Recital 67) the clarification that the requirement for complete and error-free datasets should not affect the use of privacy-enhancing techniques in the context of developing and testing AI systems. In the Parliament's later version, this clarification disappeared, but it has finally been reinstated in the final text, which we believe is positive.

As discussed in the previous section, the Council introduced a first modulation of this obligation of result, not with regard to the relevance and representativeness of the data, but with regard to the requirement of error-free and the completeness of the data, by stating that "to the greatest extent possible, they shall be error-free and complete". The Parliament, for its part, validates that the data sets are "sufficiently representative", "duly assessed for errors" and "as complete as possible in view of the intended purpose", thus clearly diluting the requirement on data quality introduced by the Parliament, which is similarly taken up in the final text. In relation to data quality, as this may depend on the context, the introduction by the Parliament of such a reference to the intended purpose of the processing is welcome. This was proposed by the Norwegian government, when it recommended including in the third paragraph a reference to the purpose of processing in the sense of relating relevance, necessity, and accuracy to the purpose of processing, as the GDPR does when defining the Data Minimisation and Accuracy Principles in Article 5.1(c) and (d) respectively.

[89] Article 43(2) and Amendment 453 as regards Article 43(1)(d).

[90] Cit. Ebers, M. *et alia*.

Regarding the *presumption*[91] that high-risk AI systems 'that have been trained and tested on data reflecting the specific geographical, behavioural, contextual or functional setting within which they are intended to be used' meet the requirements laid down in Article 10.4 is, in our view, questionable. According to this presumption, it is sufficient to 'train and test' a system with such data in order to consider that the data sets used take into account 'the characteristics or elements specific to the specific geographical, contextual, behavioural or functional environment in which the high-risk AI system is intended to be used' to the extent required for the intended purpose, which seems quite different, as the intended purpose of the system must in any case be taken into account, and these characteristics or elements will therefore vary from case to case. In any case, it is a somewhat diffuse and generic presumption to infer compliance with such an important requirement as that set out in 10.4.

In addition to the substantive content, we cannot ignore that, in order to verify compliance, in this case, with the data governance requirements, it will be necessary for the competent body to have the necessary powers to carry out *on-site* and remote unannounced inspections, as well as to access training, validation, and test data and source code of high-risk AI systems. This had been requested by the EDPS[92] and proposed by the Parliament[93]. This will require that the provider, or obliged party, is in a position to provide such samples that the national supervisory authority is empowered to request. The Parliament proposed that the obliged party should retain sufficient evidence and samples to enable the authority to "reverse engineer AI systems and acquire evidence to detect non-compliance". However, the final version[94] has not adopted this wording, but states[95] that the provider shall grant market surveillance authorities full access to the documentation, as well as to the training, validation and test data sets used and including, where appropriate and subject to security safeguards, through application programming interfaces ("APIs") or other relevant technical means and tools that allow remote access. In certain cases, access to source code will be granted[96]. It is therefore clear that the provider must retain the datasets used for the development of the system[97], which is why we believe that it would have been desirable to

[91] Article 42.1.

[92] *Cit.* EDPS, *Opinion 44/2023,* para. 45.

[93] Amendment 587 introducing a new paragraph 3a) in Article 63.

[94] Article 74.5.

[95] Article 74.12.

[96] Article 74.13.

[97] Ex Article 18, technical documentation (Annex XI) must be retained for ten years in-

clearly establish such obligation in Article 10, especially in view of the presumption discussed above, although the AIA does not expressly establish it as a *rebuttable presumption*.

As regards the *sanctioning regime* in this area, the Parliament introduced important changes. The Commission and Council versions provided for the highest penalties, on the one hand, for infringements relating to prohibited Artificial Intelligence practices (Article 5) and those relating to non-compliance with data and data governance requirements (Article 10), with fines of up to EUR 30 000 000 or, if the offender is a company, of up to 6 % of the total annual worldwide turnover in the preceding financial year, whichever is higher, and on the other hand, non-compliance with the other requirements or obligations set out in the Regulation, with administrative fines of up to EUR 20 000 000 or, if the offender is a company, up to 4 % of the total annual worldwide turnover. The Parliament proposed to increase the penalties for prohibited AI practices to EUR 40,000,000 but, interestingly, to remove from that range infringements relating to Article 10, and to create a new range of penalties for breaches of data and data governance requirements and transparency obligations[98] with penalties of EUR 20,000,000 or, if the offender is a company, up to 4% of the total annual worldwide turnover in the preceding financial year. For all other infringements of certain articles, it proposed to halve the penalties. It also proposed to halve[99] infringements for submitting inaccurate, incomplete or misleading information to notified bodies and national competent authorities, which, in a system based on "self-assessment" of compliance with requirements, is of particular relevance. Finally, the most serious penalties[100] are only for infringement of Article 5 (prohibited practices) and will entail fines of up to EUR 35,000,000 or up to 7% of its total annual worldwide turnover for the preceding business year, whichever is higher. A catalogue of certain provisions, not including Article 10, is included, the

cluding (Section 1, point 2c): "information on the data used for training, testing and validation, where appropriate, including the type and provenance of data and data management methods (e.g. cleaning, filtering, etc.), the number of data points, their scope and their main characteristics; how the data were obtained and selected, and any other measures to detect inadequacy of data sources and methods to detect biases; and any other measures to detect inadequacy of data sources and methods to detect biases.), the number of data points, their scope and their main characteristics; how the data were obtained and selected, as well as any other measures to detect inadequate data sources and methods to detect identifiable biases, where appropriate'. Note that this does not refer to the totality of the datasets.

[98] Amendment 650, Article 71, new paragraph 3a.

[99] Amendment 652.

[100] Article 99.2.

infringement of which is punishable by fines of up to EUR 15,000,000 or, if the infringer is an undertaking, up to 3% of its total annual worldwide turnover in the preceding business year. Therefore, the infringement of Article 10 has gone from being one of the most serious infringements to not appearing in the sanctioning regime, perhaps by mistake as the new sanctioning range proposed by the Parliament for infringements of Articles 10 and 13 of the AIA has been deleted from the final version.

On the other hand, the penalty for supplying incorrect, incomplete or misleading information to notified bodies and national competent authorities is increased to administrative fines of up to EUR 7,500,000 or, if the offender is an undertaking, up to 1% of its total annual turnover, so the Parliament's proposal was not accepted on this point either.

It is also striking that, despite the general mandate[101] that sanctions should take particular account of the interests of SMEs and start-ups, as well as their economic viability, the Parliament proposed to eliminate the modulation of liability introduced by the Council in relation to SMEs and start-ups, establishing a lower percentage in terms of their annual global turnover in all sanctions. The final version recovers the mention[102] to SMEs and *start-ups* and includes a modulation of liability consisting of applying the percentage or the amount of the sanction, depending on which of them is lower, contrary to what is established in the general sanctioning regime, in which the higher amount should be chosen. We consider that, although the introduction of such modulation is positive, it will only benefit those SMEs and *start-ups* whose total annual turnover is very high.

Like other authors[103], we believe that a compliance system based on "self-assessment" has been constructed without the compulsory intervention of external bodies, which, together with the reduction in penalties, even for providing inaccurate, incomplete or misleading information to the notified authorities or bodies, significantly reduces the degree of legal certainty expected to be achieved with the Regulation. Even if we think of high-risk AI systems, which have gone from being a list of *numerus clausus* to, with the amendments introduced by the Council, having to meet the cumulative criterion of posing "a significant risk to health, safety or fundamental rights",

[101] Article 99.1.

[102] Article 99.6.

[103] Cit. Ebers, M. *et alia*, p. 601; Peguera Poch, M., *La propuesta de reglamento de AI: una intervención legislativa insoslable en un contexto de incertidumbre,* Chapter closed on 20 May 2023, p. 24. Published in: Peguera Poch, Miquel (coord.) "Perspectivas regulatorias de la Inteligencia Artificial en la Unión Europea", Madrid: Reus, 2023.

it is ultimately also up to the providers to determine whether or not they are dealing with a high-risk system. EBERS *et alia*[104] sums it up nicely by stating that "in contrast to the impending over-regulation attributable to the broad definition of AI, the self-fulfilment approach raises problems of under-regulation" (translation).

V. Confluence of data protection regulation

In this section we will address the connections of data governance requirements with data protection principles, as the interaction of AIA with data protection law has been dealt with at a general level in another chapter of this work by Jiménez López.

Considering that one of the legal bases for AIA is Article 16 of the Treaty on the Functioning of the European Union (TFEU), the importance of data protection regulation in AIA is beyond doubt. It should be borne in mind that many AI systems will be trained or process personal data, or will either assist individuals in making decisions or directly be able to make and execute the decision, so the GDPR will fully apply. However, the AIA does not include within its articles a general obligation to comply with data protection regulations, without prejudice to mentions of specific obligations. The closest is the requirement[105], introduced by the Parliament and taken over in the final text, that the declaration of compliance should include a statement that the AI system complies with the GDPR.

This is not a trivial issue. Not for nothing, initially, the violation of data governance requirements was set at the same sanction level as prohibited practices. At this stage, we should not start from the premise that technology is neutral, but rather the opposite, as Floridi states[106]. Even the approach to risk regulation used is not neutral[107], so both the design and the data used are absolutely relevant, as we can see in the AIA. The consequences of not having the right type of data, nor the required quality, could be disastrous, as they condition the results from the design, thus being invalid, and more importantly, could affect the fundamental rights of individuals. The relationship between the data and the AI system is therefore directly proportional to the quality of the results obtained. This is why Article 10 includes, among good

[104] Cit. Ebers, M. *et alia*, p. 601.

[105] Annex V, point 5.

[106] Op. cit. Floridi, L., "On Good and Evil…".

[107] Op. cit. Kaminski M., "Regulating the risk of AI", p. 1351.

data governance and management practices, issues related to the design of the system and the transparency and quality of the data.

The EDPB and the EDPS stated[108] that 'the proposal (for a regulation) lacks a clear link to data protection legislation'. Other authors[109] stated that the AIA 'should aim at better harmonisation and coordination with data protection law'. This problem has been partly reduced thanks to the amendments introduced in this area by the European Parliament, by positivising in the AIA the importance of compliance with data protection rules, which, although not mentioned, is not mandatory, but highlights the importance of compliance in the field of AI.

In addition, the AIA lacks any guiding principles that would guide the different obliged parties in the application of the AIA and that would govern any interpretation by legal operators. In this regard, the Parliament proposed to introduce[110] a set of general principles applicable to all AI systems which, by informing the application of the AIA, could be enforceable on all operators within its scope, as is the case with the GDPR. However, for some unknown reason, this proposal was not taken up in the final version. Among the principles proposed by the Parliament was the principle of "*Privacy and data governance*: AI systems shall be developed and used in accordance with existing privacy and data protection rules, and shall process data that meet high standards in terms of quality and integrity". This principle is evidence of the mutual conditioning between data protection law and data governance obligations.

From the point of view of data governance obligations, possible shortcomings of the current regulation have been highlighted. However, the obligations set out in Article 10 apply irrespective of whether personal data are involved or not, and without prejudice to any obligations arising from the application of the GDPR. Likewise, allowed practices by the AIA may not be feasible if they do not comply with the requirements of data protection law[111]. It is therefore clear that data protection principles will apply in any case. However, on this point the EDPB and the EDPS[112], in relation to the

[108] EDPB-EDPS, Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council on harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act), 18 June 2021, paragraph 76.

[109] Cotino L., Castillo J.a., Salazar I., Benjamins R., Cumbreras M., Esteban A., "Un análisis crítico constructivo de la Propuesta de Reglamento de la Unión Europea por el que se establecen normas armonizadas sobre la Inteligencia Artificial (*Artificial Intelligence Act*)", in Diario La Ley, Wolters Kluwer, 2 July 2021.

[110] Amendment 213. Article 4a:

[111] Recital 63.

[112] Op. cit. EDPB-EDPS, Joint Opinion 5/2021, para. 76.

certification scheme, proposed to include the principles of minimisation and data protection by design as one of the requirements to be taken into account in order to obtain the CE marking, due to the 'possible high level of interference of high-risk AI systems with the fundamental rights to privacy and personal data protection, and the need to ensure a high level of trust in the AI system'. A view subsequently reiterated by the EDPB.[113]

Although Article 10, and the AIA in general, do not expressly state compliance with any data protection principles, the Recitals do. Thus, Recital 67 states that, in order to facilitate compliance with data protection law, data governance and management practices should include, in the case of personal data, transparency about the original purpose of data collection. Therefore, the principle of transparency in data protection becomes a condition for complying with this requirement in data governance, and vice versa, since as the AEPD states[114] "the information available under the Transparency-AIA framework should be sufficiently complete to enable controllers and processors to fulfil their different obligations under the GDPR". Recital 69 states that 'the right to privacy and to protection of personal data must be guaranteed throughout the entire lifecycle of the AI system. In this regard, the principles of data minimisation and data protection by design and by default, as set out in Union data protection law, are applicable when personal data are processed'. Recital 67 also clarifies that the requirement for data sets to be, as far as possible, complete and error-free 'should not affect the use of privacy-preserving techniques in the context of the development and testing of AI systems' and in the same vein Recital 69, where it indicates that providers to ensure compliance with these principles may use 'technology that permits algorithms to be brought to the data and allows training of AI systems without the transmission between parties or copying of the raw or structured data themselves, without prejudice to the requirements on data governance provided for in this Regulation'.

When Article 10 requires data to be error-free and complete for the intended purpose, we understand it to refer directly to the principle of accuracy. As the AEPD states[115] "the performance of an algorithm, including Artificial Intelligence (AI) algorithms, could be compromised by the inaccuracy of the

[113] EDPS, Opinion 44/2023 on the Proposal for Artificial Intelligence Act in the light of legislative developments, 23 October 2023, paragraph 27.

[114] AEPD, "Artificial Intelligence: Transparency", 20 September 2023. https://www.aepd.es/prensa-y-comunicacion/blog/inteligencia-artificial-transparencia

[115] AEPD, "Artificial Intelligence: Principle of Accuracy in Processing", 31 May 2023 https://www.aepd.es/prensa-y-comunicacion/blog/inteligencia-artificial-principio-de-exactitud-en-los-tratamientos

input data used in the execution of the algorithm, not only by the data used in its development", which is why "it is necessary to assess the accuracy of the input data, as it could introduce biases and compromise the performance not only of the algorithm, but of the entire processing".

Therefore, controls should be put in place to prevent the input of inaccurate input data and also controls to put in place adequate safeguards in case of inaccurate data input. This is the purpose of the obligation in Article 10 to put in place appropriate measures to detect, prevent and mitigate possible biases that are identified.

Precisely to ensure the detection and correction of bias in relation to high-risk AI systems, providers of such systems are exceptionally allowed to process special categories of data provided that a number of conditions are met (i) it cannot be done using synthetic, anonymised or other data; (ii) the special categories of personal data processed are subject to technical limitations on re-use and to the most advanced security measures; (iii) they are subject to measures ensuring the security and protection of the personal data processed; (iv) they are not transmitted, transferred or otherwise made accessible to third parties; (v) they are deleted once the bias has been corrected or the personal data have reached the end of their retention period. This Article refers to the principle of lawfulness for the processing of such special categories of data. Some authors[116] argue that it is an exception to the GDPR as it constitutes in itself a basis for lawfulness. On the contrary, we understand that an adequate legitimacy basis will be necessary, firstly, because of the application of the GDPR itself and, secondly, because paragraph 5 itself, when it lists the conditions necessary for the processing to take place, expressly indicates that the provisions set out in Regulation (EU) 2016/679 (…) must be taken into account.

Therefore, the importance of the principles of data protection in relation to data governance is clear, which highlights the very important interrelation and interdependence between both regulatory frameworks. So much so that if we look at both the subjective and material scope of application of the Fundamental Rights Impact Assessment[117] which has been completely blurred to the point, in our opinion, of not being able to fulfil the purpose for which it was conceived. Data protection regulations and, especially, its principles and the Data Protection Impact Assessment, ultimately stand as the guardian of the aforementioned fundamental rights, without prejudice to the fact that the risk analysis includes them within its objective scope.

[116] Op. cit. Ebers, M. *et alia*, p. 600.

[117] Article 27.

VI. Conclusions

First. At the level of AI governance, we consider that it is necessary to establish an international governance framework. Initiatives at the level of AI regulation in different continents demonstrate the need for international regulation and, therefore, the establishment of the necessary coordination mechanisms[118]. Notwithstanding the above, Europe, aware of its shortcomings in terms of technological sovereignty and seeking to safeguard health, security and fundamental rights, has established its own AI governance framework through which it aspires to repeat the "Brussels effect" it achieved with the General Data Protection Regulation. To ensure the single market for data ('macro' level governance), it is imperative that organisations have strong data governance in place internally ('micro' level), which will also enable progress towards AI governance. It should not be misconceived that these obligations only fall on the entities that develop AI systems, but that those that design or deploy them (deployers) also have responsibilities, so that, although at different levels, it is necessary for all organisations to establish AI governance mechanisms. Governance has never been more important, not only at the implementation and management level, but it must start with the management bodies that are responsible for setting and leading the AI strategy, as well as overseeing its implementation. If we had to sum up AIA in one word, it would be "Governance". In relation to data governance we consider that a broad concept should be used, not only referring to the one set out in Article 10, but also including post-marketing monitoring[119] and, in addition, long-term monitoring to detect systemic risks in relation to the gradual erosion of institutions and social and political values.[120]

Secondly, Article 10 on data and data governance is of paramount importance. Article 10 on data and data governance is of paramount importance, as compliance with the obligations set out therein results in the availability of high quality data and thus in the proper functioning of AI systems, especially high-risk ones. It sets out the requirements ('quality criteria') that datasets used for training, validation and testing of high-risk systems must meet. It is of paramount importance to have quality data for both training and system development, otherwise both the system itself and its results may be affected, which is of vital importance when we are talking about security and fundamental

[118] Roberts, H., Hine, E., Taddeo, M. and Floridi, L., "Global AI governance: barriers and pathways forward", 29 September 2023. http://dx.doi.org/10.2139/ssrn.4588040

[119] Annex IV, 2. d) and g).

[120] Op. cit. KOLT, N., Algorithmic Black Swans, p. 37.

rights. For this reason, a robust data governance system is imperative and transcendental, both to ensure the proper functioning of the system and to demonstrate the necessary proactive accountability. Data governance obligations must necessarily be connected to the quality management system and, in particular, to the risk management system, even if this is not explicitly stated. It is true that the data governance system set up by the AIA has its own entity in a way that transcends the risk management system, but this does not imply that it is unknown to the latter. The inclusion of the express reference to Article 10 would have been desirable, both in the quality management system and in the risk analysis, not only for the sake of emphasising the importance of compliance with Article 10, but also for reasons of systematic consistency.

Third. We understand that all the circumstances are ripe for the emergence in the value chain of new figures that exclusively provide "verified" data and certify that the data comply with the established governance requirements, as well as their integrity and training ("data verifiers" or "certified data service providers"). Regulation and possible transfer of responsibility will therefore be key. We have questioned whether this model of "verified" data provision can deliver the individualised compliance with such governance requirements to which the Regulation aspires since, firstly, governance must be tailored to the context of use as well as the intended purpose of the AI system[121] and, secondly, data sets should take into account, to the extent required by the intended purpose, the characteristics or elements specific to the geographical, behavioural contextual or functional environment in which the high-risk AI system is intended to be used. It will therefore be the ultimate responsibility of the deployer to assess the appropriateness of such datasets for the use case for which the AI system will be used. In other words, the fact that new figures may enter the value chain as a result of the outsourcing of data governance requirements does not exempt the deployer (and, where applicable, data controller) from compliance with the other obligations, as "proactive accountability is a cornerstone of AI governance"[122] both proactive (*ex ante*) and reactive (*ex post*). Without prejudice to the questioning of this model, the regulation and possible transfer of liability at the contractual level will be key.

Fourth. In order to verify compliance with the data governance requirements by the competent authorities, the provider or obliged party shall retain the data sets used for the development and training of the system. This follows from the post-market surveillance measures[123] stating that "providers

[121] Article 10, second paragraph.

[122] Op. cit. Novelli C., Taddeo M., Floridi L., Accountability in Artificial Intelligence.

[123] Article 74(12).

shall grant market surveillance authorities full access to documentation as well as training, validation and test data sets used for the development of high-risk AI systems". For this reason, we believe that it would have been desirable to clearly state this obligation in Article 10 itself. Furthermore, it should be evidenced that the AI system complies with the GDPR, with all that this implies, as stated in the declaration of conformity.

Fifth. Although not expressly included, the application of data protection law will play an absolutely necessary role as regards the quality criteria to be met by data sets, as data protection principles must be observed. In addition, data protection law, through its Data Protection Impact Assessment, will play a fundamental role in safeguarding the rights and freedoms of data subjects, partly filling the gap that should be filled by the Data Protection Impact Assessment.

Sixth. AIA is a compliance challenge, as it brings together a complex set of technical and/or harmonised requirements and standards, together with the application of data protection and fundamental rights regulations, to which we must add the interaction of different roles whose compliance must ultimately be supervised by the deployer. While it is true that the nature of AI makes it problematic to assess accountability for the results obtained[124], this shifts the burden of proactive accountability to being able to demonstrate relevance, sufficient representativeness, error analysis, and completeness of data, which will be achieved with robust data governance systems.

[124] Op. cit. Novelli C., Taddeo M., Floridi L., Accountability in artifcial intelligence.

QUALITY MANAGEMENT SYSTEMS, TECHNICAL DOCUMENTATION AND DOCUMENTATION KEEPING IN THE REGULATION

Francisca Ramón Fernández
Professor of Civil Law, Universitat Politècnica de València[1]

I. Introduction

In this study we are going to deal with the regulation of quality management systems, technical documentation and preservation of high-impact Artificial Intelligence systems in the different proposals, as well as in the final text of the Artificial Intelligence Act in relation to high-risk Artificial Intelligence systems covered by the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union COM(2021) 206 final 2021/0106 (COD), of 21 April 2021[2], together with Annexes[3], and in the final text P9_TA(2024)0138 on the European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union (COM(2021)0206 - C9-0146/2021 -2021/0106(COD)), AIA.[4]

The regulation under analysis is contained in Chapter (in previous versions referred to as Title) III dedicated to regulating high-risk Artificial In-

[1] Work carried out in the framework of the Generalitat Valenciana Research Group of Excellence "Algorithmical Law" (Project Prometeu 2021/009, 2021-2024), and Project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by MCIN/AEI/10.13039/501100011033/ ERDF, EU.

[2] Text of the Proposal available at: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF (Accessed 25 July 2023).

[3] Available at: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_2&format=PDF (Accessed 25 July 2023).

[4] Text of the Artificial Intelligence Regulation available at: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_ES.pdf (Accessed 15 March 2024). See also: Mühlhoff, R. and Ruschemeier, H., "Regulating AI with purpose limitation for models", *Journal of AI Law and Regulation*, No. 1 (2024), pp. 24-39. Available at: https://aire.lexxion.eu/data/article/19395/pdf/aire_2024_01-006.pdf (Accessed 18 March 2024).

telligence systems, in Section (previously in previous versions referred to as Chapter) 2 dedicated to requirements for high-risk Artificial Intelligence systems, in Article 11 which refers to technical documentation. These requirements are derived from the ethical guidelines for trustworthy Artificial Intelligence that were developed by the independent high-level expert group on Artificial Intelligence that was set up by the European Commission, in June 2018[5]. Flexibility is provided for with regard to the technical solutions needed to achieve compliance with the indicated requirements which may be derived from technical standards or specifications or be subject to development in accordance with scientific or engineering knowledge, on a discretionary basis by the AI system provider. This allows system providers to decide how they want to meet the requirements, taking into account the state of the art and technological and scientific developments.

In section (previously referred to as Chapter in previous versions) 3 dealing with the obligations of providers and deployers (formerly users) of high-risk Artificial Intelligence systems and other parties, in particular Article 17 on the quality management system, and Article 18 initially intended to cover the obligation to produce technical documentation, and then in the subsequent version of the 2022 Proposal for a Regulation and the 2024 European Parliament Legislative Resolution, renamed the document retention, and to include the content of Article 50 contained in section (previously referred to as chapter 5 in previous versions) 5 on standards, conformity assessment, certificates, registration and which dealt with the retention of documents, with this Article 50 being subsequently deleted.

It is a set of horizontal obligations imposed on providers of high-risk Artificial Intelligence systems[6], and also places obligations on users and other participants in the Artificial Intelligence value chain, such as importers, distributors and authorised representatives.[7]

[5] European Union, *Ethical Guidelines for Trustworthy AI. High Level Expert Group on Artificial Intelligence*, European Commission, Brussels 2019.

[6] See: Cotino Hueso, L., "Los usos de la inteligencia artificial en el sector público, su variable impacto y categorización jurídica", *Revista Canaria de Administración Pública*, n.º 1 (2023), pp. 211-242. Available at: https://revistacanarias.tirant.com/index.php/revista-canaria/article/view/7/7 (Accessed 24 July 2023).

[7] Cfr. Ramón Fernández, F., "Inteligencia artificial y transparencia en relación con la regulación de los servicios y mercados digitales", *Equidad y transparencia en la prestación de servicios*, María Elena Cobas Cobiella and Raquel Guillén Catalán, editors, Dykinson, Madrid (2023), pp. 147-169. Also of interest: Argelich Comelles, C., "Gobernanza de las plataformas en línea ante la DSA y las Propuestas de Reglamento de Mercados Digitales e Inteligencia Artificial (DMA y AIA). (Gobernanza de plataformas en línea frente a DSA, DMA y AIA de la UE)", *Anuario de*

As stated in Recital 9, "this Regulation aims to strengthen the effectiveness of such existing rights and remedies by establishing specific requirements and obligations, including in respect of the transparency, technical documentation and record-keeping of AI systems". Similarly, Recital 66 states that "Requirements should apply to high-risk AI systems as regards risk management, the quality and relevance of data sets used, technical documentation and record-keeping, transparency and the provision of information to deployers, human oversight, and robustness, accuracy and cybersecurity. Those requirements are necessary to effectively mitigate the risks for health, safety and fundamental rights. As no other less trade restrictive measures are reasonably available those requirements are not unjustified restrictions to trade".

In the former Chapter 5, where Article 50 was located, the conformity assessment procedures to be followed for each type of high risk AI system were explained in detail. This was intended to reduce the burden on economic operators and notified bodies. AI systems intended to be used as safety components of products covered by the legislation of the new regulatory framework, e.g., machines, medical devices or toys, will be subject to the same compliance mechanisms upstream and downstream as the products in which they are integrated.[8]

Derecho Civil, vol. II, pp. 501-530. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4434522 (Accessed 11 November 2023).

[8] As indicated in the 2021 Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of AI (Artificial Intelligence Act) and amending certain Union legislation, "a new compliance and enforcement system will be established for the stand-alone high-risk AI systems referred to in Annex III. This will be modelled on the NRA legislation and implemented by providers through internal controls, with the exception of remote biometric identification systems, which will be subject to third-party compliance assessments. An effective and reasonable solution for such systems could be a comprehensive ex-ante conformity assessment through internal controls, combined with stringent ex-post monitoring, given that regulatory intervention is at an early stage, that the AI sector is highly innovative and that the necessary expertise to conduct audits is just beginning to accumulate. In order to assess "stand-alone" high-risk AI systems through internal controls, full, effective and properly documented ex-ante compliance with all the requirements of the Regulation, as well as robust quality and risk management systems and post-market monitoring, would be necessary. Once the provider has carried out timely conformity assessment, it should register such independent high-risk AI systems in an EU database to be managed by the Commission for the purpose of enhancing public transparency and vigilance and strengthening ex-post monitoring by competent authorities. Instead, for reasons of consistency with existing product safety legislation, the conformity assessment of AI systems that are product safety components will follow a system where third parties will carry out conformity assessment procedures already defined in the relevant sectoral product safety legislation. If sub-

The provisions of Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery should be taken into account[9]. Highlight the importance of the regulation under study. Artificial Intelligence is also one of the five interrelated specific objectives of the Digital Europe Programme set out in Regulation (EU) 2021/694 of the European Parliament and of the Council of 29 April 2021 and repealing Decision (EU) 2015/2240.[10]

Title III contains specific rules for Artificial Intelligence systems that pose a high risk to the health and safety or fundamental rights of individuals. The risks that may arise from the implementation of these systems on the European market must be weighed up, provided that they comply with mandatory requirements and are assessed prior to their introduction on the EU market. The function, purpose, and modalities of use of the system will be the factors for the qualification of a high-risk Artificial Intelligence system.

The Report from the Commission to the European Parliament, the Council and the European Economic and Social Committee on the security and liability implications of Artificial Intelligence, the internet of things and robotics annexed to the White Paper on Artificial Intelligence - A European approach to excellence and trust, of 19 February 2020 [COM(2020) 65 final][11] indicates that "the autonomous behaviour of some AI systems throughout their lifecycle may lead to significant product changes and safety implications, which may require a new risk assessment. In addition, human supervision is likely to be required as a safeguard from the design phase and throughout the life cycle of AI products and systems".[12]

Regarding this high risk and the need for ex ante control and assessment, it is worth mentioning the White Paper on Artificial Intelligence - A European approach to excellence and trust COM(2020) 65 final of 19 February 2020.[13] On an objective ex-ante compliance check to verify and ensure com-

stantial modifications are made to the AI systems (mainly changes that go beyond the aspects pre-determined by the provider in his documentation".

[9] Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0132_ES-.pdf (Accessed 15 March 2024).

[10] OJEU L 166 of 11 May 2021. Available at: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32021R0694 (Accessed 24 July 2023).

[11] Available at: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX-:52020DC0064 (Accessed 25 July 2023).

[12] Ramón Fernández, F., "El robot como producto defectuoso y responsabilidad civil", *Derecho Digital e Innovación*, n.º 14 (2022), pp. 1-28.

[13] Available at: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX-:52020DC0065 (Accessed 25 July 2023).

pliance of high-risk applications with some of the above-mentioned mandatory requirements, which may include testing, inspection, or certification procedures, as well as having checks on the algorithms and datasets used in the development phase.

We refer to Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 ("Cybersecurity Regulation"). It is also worth noting the Commission Decision of 24 January 2024 establishing the European Office for Artificial Intelligence.

We should also mention, in the Spanish sphere, Royal Decree 817/2023 of 8 November, which establishes a sandbox (sandbox) for testing compliance with the proposed Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence, which states that "the first sandbox is being set up to test how to implement the requirements applicable to high-risk Artificial Intelligence systems of the proposed European regulation on Artificial Intelligence with the aim of obtaining, as a result of this experience, guidelines based on evidence and experimentation that will help entities, especially small and medium-sized enterprises, and society in general, to align with the proposed European Regulation on Artificial Intelligence, as a result of this experience, evidence- and experimentation-based guidelines that will facilitate the alignment of organisations, especially small and medium-sized enterprises, and society in general, with the proposed European Artificial Intelligence Regulation. During the development of this sandbox, the position of the Council of the European Union of 25 November 2022 will be used as a reference".[14]

The main issues to be addressed will be the following:

a) The different changes and modifications made in the successive versions proposed after the initial text of 2021 will be analysed, in relation to the final text approved in 2024, with the aim of observing what has been their purpose and aim with regard to Articles 11, 17, 18 and 50 (currently deleted

[14] Royal Decree 817/2023 also states, "Artificial Intelligence is a disruptive technology with a high capacity to impact the economy and society. In economic terms, and together with other digital technologies, it has a high potential for increasing productivity, opening up new lines of business, developing new products or services - based, for example, on personalisation, optimisation of industrial processes or value chains -, improving the ease of performing everyday tasks, automating certain routine tasks and developing innovation. This potential has a positive impact on economic growth, job creation and social progress.

However, Artificial Intelligence systems may also pose risks to the respect of citizens' fundamental rights, such as those relating to discrimination and the protection of personal data, or even cause serious problems for the health or safety of citizens.

from the initial content and which now deals with regulating Transparency Obligations of providers and users of certain AI systems).

b) Identify the main reasons for the changes made, as well as the most relevant aspects of their implementation.

c) Establish the contexts where high-risk Artificial Intelligence systems may operate and how to establish the quality management system, technical documentation issues and document retention aspects.

The methodology we are going to use is to carry out a comparative analysis of the different regulations applicable to AI according to the different versions of the proposals, as well as the doctrine that has been pronounced on the matter in order to obtain valid conclusions applicable to the international scientific community.

II. Article 17 of the AIA on the quality management system

In the first text prepared by the European Commission in 2021, it was established that providers of high-risk AI systems shall establish a quality management system, and that it shall be documented in a systematic and orderly manner through written policies, procedures, and instructions that shall include aspects specified in the precept itself (techniques, procedures, examination, testing and validation, technical specifications, data management, risk management, notifications, registration, accountability, among others), and that it shall be proportional to the size of the provider's organisation.

Regarding data management, the Proposal for a Regulation of the European Parliament and of the Council on harmonised rules for fair access to and use of data (Data Law) of 23 February 2022 [COM(2022) 68 final 2022/0047 (COD)][15], should be taken into account, and also, the European Parliament legislative resolution of 9 November 2023 on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act) [COM(2022)0068 - C9-0051/2022 - 2022/0047(COD)].[16]

In the case of providers that are credit institutions covered by Directive 2013/36/EU of the European Parliament and of the Council of 26 June 2013 relating to the taking up and pursuit of the business of credit institutions and the prudential supervision of credit institutions and investment

[15] Available at: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX-:52022PC0068 (Accessed 13 November 20223).

[16] Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0385_EN.pdf (Accessed 13 November 2023).

firms, amending Directive 2002/87/EC and repealing Directives 2006/48/EC and 2006/49/EC, shall be deemed to comply with the obligation to establish a quality management system where they meet the standards for governance systems, procedures and arrangements referred to in Article 74 of that Directive. In this context, account shall be taken of all harmonised standards referred to in Article 40 of the AIA, which mentions Regulation (EU) No. 1025/2012 of the European Parliament and of the Council of the European Union. 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council.

The AIA Proposal introduces a new paragraph 2a in order to ensure greater harmonisation with sectoral legislation on obligations related to quality management systems.

This new paragraph indicates that for providers of high-risk Artificial Intelligence systems that are subject to obligations relating to quality management systems under relevant sectoral Union law, the aspects described in paragraph 1 may form part of the quality management systems under that law.

In Article 17.3 of the AIA in the 2022 Proposal, the Fourth Presidency Compromise Text, several adjustments are made by mentioning only financial institutions without specifying credit institutions, as well as the deletion of the reference to Directive 2013/36/EU, and the reference to the derogation in paragraph 1(g), (h) and (i) and the reference to the relevant Union financial services legislation.

The version incorporating the amendments adopted by the European Parliament is the text called the Artificial Intelligence Act Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union (COM(2021)0206 - C9-0146/2021 - 2021/0106(COD)).[17]

Amendment 346 on the proposed Regulation, with regard to Article 17.1,

[17] Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_ES-.pdf (Accessed 24 July 2023). See also: BARRIO ANDRÉS, M., "Novedades en la tramitación del próximo Reglamento europeo de inteligencia artificial", *Real Instituto Elcano*, (2023), pp.

introductory part of the Commission text, makes a difference to the wording in that providers of high-risk AI systems no longer indicate that they shall establish, which implied an obligation, but is replaced by 'shall provide' as a power of disposal of the quality management system. It also introduces an alternative that was not in the initial wording, as it now refers to written procedures or instructions, and adds the possibility of introducing into an existing quality management system in accordance with sectoral Union legislation. The previous version of the adopted text stated that management systems shall be implemented, thus reverting to the mandatory nature of the quality management system, and no longer mentions procedures or written instructions as an alternative, but both procedures and written instructions, as the wording has been changed to include both. However, the final text reverts to the original wording and reads as follows: "Providers of high-risk AI systems shall put a quality management system in place that ensures compliance with this Regulation. That system shall be documentes in a systematic and orderly manner in the form of written policies, procedures and instructions'.

In amendment 347 on Article 17.1(a), the indication to include "(a) a strategy for regulatory compliance, including compliance with conformity assessment procedures and procedures for managing changes to high-risk AI systems" was deleted.

The adopted text reinstates the previously deleted paragraph 1(a) and states that it shall include "(a) a strategy for regulatory compliance, including compliance with conformity assessment procedures and procedures for the management of modifications to the high-risk AI system". Clearly the reference to regulatory compliance and not merely regulatory compliance is expanded.

In amendment 348, with regard to Article 17.1(e) dealing with the technical specifications, including standards, to be applied and, where the relevant harmonised standards are not applied in full, adding "do not cover all relevant requirements", the means to be used to ensure that the high risk AI system complies with the requirements set out in Chapter 2 of this Title.

The adopted text includes the wording of amendment 348 so that it refers to the technical specifications, including standards, to be applied and, where the relevant harmonised standards are not applied in full or do not cover all the relevant requirements laid down in Chapter II, the means to be used to ensure that the high-risk AI system complies with the requirements laid down by referring to the above-mentioned Chapter. The reformulation

1-10. Available at: https://www.realinstitutoelcano.org/analisis/novedades-en-la-tramitacion-del-proximo-reglamento-europeo-de-inteligencia-artificial/ (Accessed 24 July 2023).

of the initial wording makes the provision more technically agile and easier to understand, with the aim of avoiding legislative repetition.

Article 17.1(f) is amended by including in respect of data management systems and procedures, including data acquisition, data collection, data analysis, data labelling, data storage, data filtration, data mining, data aggregation, data retention and any other operation regarding the data that is performed before and for the purpose of the placing on the market or the putting into service of high-risk AI systems. The adopted text maintains the reference to data acquisition and data collection already mentioned in the amendment.

In amendment 350, concerning Article 17.1(j), the reference to competent national authorities is deleted, indicating only that the management of the communication shall be carried out with the relevant national authorities, including sectoral authorities. Rather than a competence aspect, it is considered an aspect of adequacy or relevance, and the indication initially contained that they allow access to data or facilitate access to data is deleted; and also the reference to notified bodies; other operators; clients; or other interested parties. The adopted text clarifies by indicating national authorities, deleting the indication of competent, other relevant competent authorities, including sectoral competent authorities.

In the proposed amendment 351 to Article 17.2, concerning the inclusion of the aspects mentioned in paragraph 1 which shall be proportionate to the size of the provider's organisation, a paragraph is added stating that "providers shall, in any event, respect the degree of rigour and the level of protection required to ensure the compliance of their high-risk AI systems with this Regulation".

This amendment is maintained in the adopted text.

With regard to the application of quality management, the doctrine has indicated a case[18]. This is the case of quality management applied to banknote production. Thus, it has been suggested that by applying the concept of quality 4.0, which encompasses various aspects in which Industry 4.0 enabling technologies can improve product quality management systems. The tools to be implemented would be improved connectivity, data analysis, Artificial Intelligence and automation.

This author highlights several points to be taken into account[19]: a) Edge Computing and IoT (Internet of Things) networks provide a greater volume

[18] I follow the discussion in López González, A., "Inteligencia artificial aplicada al control de calidad en la producción de billetes", *Papel ocasional del Banco de España*, No. 2303 (2023), pp. 1 ff. Available at: https://ssrn.com/abstract=4451046 (Accessed on 11 November 2023).

[19] *Ibid*, pp. 12 and 13.

of reliable data for analysis; b) Data analysis allows for more accurate data-driven decision making and modelling for event prediction and production forecasting; c) 5G and other connectivity improvements enhace the speed of information exchange both within and outside the environment; d) Collaboration and compliance between the different agents linked to production are achieved through the intranet and document management, and are combined with *blockchain* networks *that* consolidate the security of information exchange; and e) The promotion of the culture of quality at all levels of the company or entity is key to ensuring compliance by all the links involved in the chain.

Article 17.2(a) is introduced by Council Mandate and the same wording is retained in the draft text stating that: "2a. For providers of high-risk AI systems which are subject to obligations relating to quality management systems or their equivalent function under relevant sectoral Union legislation, the aspects described in paragraph 1 may form part of the quality management systems under such legislation".

This paragraph becomes Article 17.3 with some interesting modifications in its final wording concerning the applicable regulation not limited to legislation but to law: "Providers of high-risk AI systems that are subject to obligations regarding quality management systems or an equivalent function under relevant sectoral Union law may include the aspects listed in paragraph 1 as part of the quality management systems pursuant to that law".

Article 17.4 (previously paragraph 3) maintains the wording of the text of the Draft Council Mandate with regard to providers that are financial credit institutions to which Directive 2013/36/EU applies, are required to establish quality management with the exception of paragraph 1, (g), (h) and (i) shall be deemed to be compliant if the rules on internal governance arrangements, mechanisms or processes in accordance with relevant Union financial services legislation are respected. Taking into account the harmonised standards referred to in Article 40.

Article 16 concerning the obligations of providers of high-risk AI systems refers to Article 17, as these providers must have a quality management system in line with the above-mentioned provision.

Article 63 on derogations for specific operators and in respect of microenterprises as defined in Commission Recommendation 2003/361/EC provided that they do not have associated or related undertakings may comply with certain elements of the quality management system required by Article 17 of the AIA in a simplified form. To this end, the Commission will develop guidelines on the elements of the system that can be complied with in such a way without affecting the level of protection and the need to comply with the requirements of high-risk Artificial Intelligence systems.

Annex VI refers to Article 17 on conformity assessment procedure based on internal control, as the provider shall verify that the quality management system in place complies with the requirements of that provision. Also Annex VII on conformity based on assessment of the quality management system and assessment of the technical documentation referred to in Article 17. This quality management system will be evaluated by the notified body who will determine if it covers all aspects referred to in Article 17.

III. Article 11 of the AIA on technical documentation with Annexes

The 2022 Proposal for a Regulation of the European Parliament and of the Council incorporated some differences with the initial 2021 text, as it established with regard to technical documentation in high-risk Artificial Intelligence systems that all information must be provided to national competent authorities and notified bodies in a clear and complete manner, which was not contained in the 2021 wording. SMEs are included and mention is made of start-ups, which was not specified in the 2021 wording, and in this case would contain, as a minimum, any documentation equivalent to the elements set out in Annex IV that meets the same objectives, unless deemed inadequate, and the indication of prior approval by the competent authority is deleted.

The version of the Artificial Intelligence Act from June 2023 incorporates Amendment 292 with regard to Article 11.1, which originally stated: "The technical documentation shall be written in such a way as to demonstrate that the high-risk AI system complies with the requirements set out in this Chapter and shall provide national competent authorities and notified bodies with *all* the information they need to assess whether the AI system concerned complies with those requirements. It shall contain at least the elements set out in Annex IV" and is replaced by "The technical documentation shall be drawn up in such a way as to demonstrate that the high-risk AI system complies with the requirements set out in this Section and to provide national competent authorities and notified bodies with the necessary information in a clear and comprehensive form to assess the compliance of the AI system with those requirements. It shall contain, at a minimum, the elements set out in Annex IV. SMEs, including start-ups, may provide the elements of the technical documentation specified in Annex IV in a simplified manner".

The reference to small and medium-sized enterprises and start-ups stands out in this amendment, so that it will be necessary to take into account the Commission Regulation (EU) No. 651/2014 of 17 June 2014 declaring certain categories of aid compatible with the internal market in application of

Articles 107 and 108 of the Treaty, which considers an enterprise, according to Article 1 of Annex I, to be "any entity", regardless of its legal form, which carries out an economic activity. In particular, entities engaged in a craft activity or other activities on an individual or family basis, as well as partnerships and associations engaged in a regular economic activity, are considered to be undertakings. As regards what is considered to be an SME, Article 2 of Annex I includes small enterprises which employ fewer than 50 persons and whose annual turnover or annual balance sheet total does not exceed EUR 10 million.

Similarly, the specific mention of SMEs in the amendment may have its basis in Regulation (EU) 2021/694 which, in its Article 5, focusing on the objective of Artificial Intelligence, pursues as an operational objective to make development and enhancement capabilities and basic knowledge of Artificial Intelligence accessible to enterprises, and in particular to SMEs and start-ups.

The text adopted maintains the wording of the Council Mandate, and adds in the case of SMEs, including newly created ones using the text the denomination "emerging" which may provide the elements of the technical documentation specified in Annex IV in a simplified form. To this end, the Commission shall establish a simplified technical documentation form geared to the needs of small and micro-enterprises. Where an existing or newly established SME, in terms of the "emerging" text, chooses to provide the information required in Annex IV in a simplified form, it shall use the form referred to in the precept. Notified bodies shall accept the form for conformity assessment purposes.

In the text we also note different references to Article 11 as is the case of Article 22 which refers to authorised representatives of providers of high-risk AI systems in which before a high-risk AI system is placed on the market, they shall ensure that it complies with the Regulation and verify that the EU declaration of conformity and the technical documentation referred to in Article 11 have been drawn up and that the provider has carried out an appropriate conformity assessment procedure. The reference to Annex IV which refers to the technical documentation referred to in Article 11.1 of the AIA is deleted and shall contain minimum information on the relevant Artificial Intelligence system such as a detailed description of the elements and the process for its development, as well as detailed information on the supervision, operation and control of the Artificial Intelligence system. Within each of these blocks is a detailed specification of the system version, software, interface, system logic and algorithms, data requirements in datasheets, validation and testing procedures, cybersecurity measures, among others.

Article 11 is also referred to in Article 97 on the exercise of delegation with regard to the delegation of powers under Chapter XI concerning the delegation of power and committee procedure.

It is also necessary to take into account the provisions of Spanish Law 29/2022 of 21 December on the promotion of the start-up ecosystem and Law 18/2022 of 28 September on the creation and growth of companies, as well as Decree-Law 2/2023 of 8 March on urgent measures to promote Artificial Intelligence in Extremadura and Royal Decree 729/2023 of 22 August approving the Statute of the Spanish Artificial Intelligence Oversight Agency.

All this in accordance with the document prepared by the Government of Spain *Digital Spain 2026*[20] whose purpose for that date is to accelerate the digitalisation of companies, focusing mainly on SMEs and start-ups, and to create favourable conditions for the emergence and maturation of technology-based start-ups, following the indications of the National Artificial Intelligence Strategy (ENIA)[21], which is one of the components of the Recovery, Transformation and Resilience Plan.[22] It is part of axis 6 of the Strategy, which corresponds to Component 16 in the Recovery Plan, and one of its objectives is to support the mass deployment and use of Artificial Intelligence by large companies, public administrations, small and medium-sized enterprises, start-ups and civil society.

It is worth mentioning the aforementioned Royal Decree 817/2023 which, in accordance with the provisions of Law 28/2022, of 21 December, on the promotion of the start-up ecosystem[23], in Article 16, provides for the creation of controlled environments, for limited periods of time, in order

[20] Available at: https://espanadigital.gob.es/sites/espanadigital/files/2022-07/EspañaDigital_2026.pdf (Accessed 24 July 2023). Previously, it is worth noting the document also produced by the Spanish Government called *España Digital 2025*. Available at: https://avancedigital.mineco.gob.es/programas-avance-digital/Documents/EspanaDigital_2025_TransicionDigital.pdf (Accessed 6 November 2023), which already mentioned the acceleration of the digitalisation of companies, with special attention to micro-SMEs, as well as the Digital Agenda Strategy 2021-2027 (https://www.europarl.europa.eu/factsheets/es/sheet/64/una-agenda-digital-para-europa, accessed 6 November 2023), which addresses connectivity, technology infrastructures, digital talent and the digital economy. See also the *Digital Bill of Rights*, 2021. Available at: https://www.lamoncloa.gob.es/presidente/actividades/Documents/2021/140721-Carta_Derechos_Digitales_RedEs.pdf (Accessed 6 November 2023).

[21] Available at: https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/ENIA2B.pdf (Accessed 24 July 2023).

[22] Available at: https://portal.mineco.gob.es/es-es/ministerio/plan_recuperacion/Documents/Plan-de-Recuperacion-Transformacion-Resiliencia.pdf (Accessed 24 July 2023).

[23] It is also worth mentioning Order PCM/825/2023, of 20 July, which regulates the criteria and procedure for the certification of start-ups that give access to the benefits and specialities recognised in Law 28/2022, of 21 December, on the promotion of the start-up ecosystem (BOE no. 173, of 21 July 2023).

to assess the usefulness, viability, and impact of technological innovations applied to regulated activities, to the supply or provision of new goods or services, to new forms of supply or provision thereof or to alternative formulas for their supervision or control by the competent authorities. The creation of sandboxes for the assessment of their impact is indicated as being justified by overriding reasons of general interest.

This Royal Decree 817/2023, as stated in Article 1, "aims to establish a sandbox to test compliance with certain requirements by some Artificial Intelligence systems that may pose risks to the safety, health and fundamental rights of individuals. It also regulates the procedure for the selection of the systems and entities that will participate in the sandbox".

It mentions quality management in the reference to "self-assessment of compliance" in Article 3, considered as the procedure for verifying compliance with the requirements, the quality management system, the technical documentation and the post-market surveillance plan. Article 13 states that both the participating AI provider and, where applicable, the participating user shall carry out the following actions to complete the self-assessment, one of these actions being the verification that the quality management system complies with the specifications provided by the competent body. The competent body shall examine the documents associated with the declaration of compliance submitted by the AI provider, mainly those describing the quality management system, the technical documentation or the post-market surveillance plan, as also provided for in Article 13 of Royal Decree 817/2023.

In Article 11, a paragraph 3a (new) was introduced in amendment 294 compared to the Commission text. This new paragraph provides that "In the case of providers which are credit institutions covered by Directive 2013/36/EU, the technical documentation shall form part of the documentation relating to the systems, procedures and internal governance mechanisms set out in Article 74 of that Directive".

This new paragraph incorporates what was indicated in Article 18.2, which is deleted by Amendment 354.

Recital 158 of the final text specifies the following: "To further enhance the consistency between this Regulation and the rules applicable to credit institutions regulated under Directive 2013/36/EU, it is also appropriate to integrate some of the providers' procedural obligations in relation to risk management, post marketing monitoring and documentation into the existing obligations and procedures under Directive 2013/36/EU. In order to avoid overlaps, limited derogations should also be envisaged in relation to the quality management system of providers and the monitoring obligation placed on deployers of high-risk AI systems to the extent that these apply to credit institutions regulated by Directive 2013/36/EU."

It refers to institutions' systems, procedures, and mechanisms, focusing on internal governance and recovery and resolution plans, with mention of the European Banking Authority (EBA).

In amendment 293 regarding Article 11.2, the wording is changed to the effect that when a high-risk AI system associated with a product covered by the legislative acts referred to in Annex II, Section A is placed on the market or put into service, a single technical documentation shall be drawn up containing all the information stipulated in Annex 1, instead of Annex IV as originally drafted, as well as the information required by those legislative acts.

Royal Decree 817/2023, in its Annex VI, mentions the technical documentation to be submitted upon completion of the implementation of the requirements, to which reference should be made. All information provided will be treated with due confidentiality in accordance with Article 18 of this Royal Decree.

As indicated in Article 11 of Royal Decree 817/2023, this technical documentation of the Artificial Intelligence system listed in Annex VI shall be prepared in accordance with specifications to be provided by the competent body and shall be updated throughout the duration of the sandbox.

According to Article 13 of Royal Decree 817/2023, both the participating AI provider and, if applicable, the participating user must carry out the following actions to complete the self-assessment, including verifying that the design and development of the Artificial Intelligence system process and its post-marketing monitoring are consistent with the technical documentation and the specifications provided by the competent body, and it shall also verify that the technical documentation of its Artificial Intelligence system includes the content according to the specifications of Annex VI of the aforementioned Royal Decree, in addition to the documentation verifying compliance with the above points indicated in Article 13 above.

As indicated in Article 14 of Royal Decree 817/2023, with regard to post-marketing monitoring, it will be based on a post-marketing monitoring plan to be included in the technical documentation to be provided, which is included in Annex VI of this Royal Decree. For its drafting, the specifications provided by the competent body for this purpose shall be followed.

Article 21 of Royal Decree 817/2023 with regard to obtaining information on the development of the environment states that during the course of the sandbox, the General Subdirectorate for Artificial Intelligence and Digital Enabling Technologies shall collect information from both the participating AI providers and the participating users on how the relevant actions have been implemented in each Artificial Intelligence system; how the self-assessment of compliance has been carried out; the technical documentation associated with each AI system; and on the quality or risk management systems described in the annexes or guides.

IV. Article 18 of the AIA on documentation keeping

The 2022 Proposal for a Regulation of the European Parliament and of the Council leaves Article 18 in its initial wording on the obligation to draw up technical documentation without content, and takes over the content of Article 50 on the retention of documents, which is referred to as "Document retention". It updates Article 18.2 in line with the changes introduced in the third compromise text in relation to financial institutions and reflects these changes in Article 20.2.

Article 18, which as drafted by the proposed regulation in paragraph 1 indicated that providers of high-risk AI systems shall draw up the technical documentation referred to in Article 11 in accordance with Annex IV, is deleted by amendment 353, and Article 18.2 is also deleted, by amendment 354, the wording of which provided that in the case of providers which are credit institutions covered by Directive 2013/36/EU, the technical documentation shall be part of the documentation relating to the systems, procedures and internal governance arrangements set out in Article 74 of that Directive.

The adopted text maintains the wording of the Council's mandate.

Article 16 with regard to the obligations of providers of high-risk Artificial Intelligence systems mentions as one of them the retention of the documentation referred to in Article 18.

V. The initial Article 50 of the AIA on document retention

Article 50[24] referring to "Document retention" in the 2022 Proposal for a

[24] This precept in the wording of the Proposal for a Regulation provided that:

"The provider shall, for a period ending 10 years after the AI system has been placed on the market or put into service, keep at the disposal of the national competent authorities:

(a) the technical documentation referred to in Article 11;

(b) the documentation concerning the quality management system referred to Article 17;

(c) the documentation concerning the changes approved by notified bodies where applicable;

(d) the decisions and other documents issued by notified bodies where applicable;

(e) the EU declaration of conformity referred to in Article 48.

Regulation is eliminated and its content is reproduced in Article 18, which is left without the initial content referring to the obligation to prepare technical documentation, and is renamed "Documentation keeping".[25]

Amendment 477 on the proposed Regulation added to the first paragraph of Article 50 the indication of the national supervisory authority in addition to the national competent authorities for a period ending ten years after the Artificial Intelligence system has been placed on the market or put into service.

In the approved text, the elimination of the provision is maintained, although the numbering of the provision is recovered within Chapter IV "Transparency obligations for providers and those responsible for the deployment of certain AI systems" under the title "Transparency obligations for providers and deployers of certain AI systems".

VI. Conclusions

In this paper we have analysed the quality management, the technical documentation and retention systems, in particular articles 17, 11, 18, and 50 of the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union, and in comparison with the final text of the European Parliament's legislative resolution of 13 March 2024, pointing out the most relevant changes that have been made throughout the legislative trajectory. Changes can be seen in the numbering of the articles and also in the terminology used, with the reference to "deployers" when previously it referred to "users", and the structure in chapters and sections (previously to title and chapter), among others.

Through the analysis of the text of the Commission's proposal and the amendments agreed by the Parliament, in which some changes are made to the initial wording, we have observed some aspects that may be of interest. In addition, the recent publication of Royal Decree 817/2023 should be taken

[25] See: Zapata Cárdenas, C. A. and Giménez Chornet, V., "Retos de los archivos ante los derechos digitales", *Los nuevos retos de los Derechos Digitales*, Ramón Fernández, F. (coord.), Tirant lo Blanch, Valencia, 2022, pp. 313 ff; and Giménez Chornet, V., "La problemática de la inteligencia artificial en gestión documental documentalística", Ciencia de Datos y Perspectivas de Inteligencia Artificial, Ramón Fernández, F. (coord.), Tirant lo Blanch, Valencia, 2022, pp. 313 ff, "La problemática de la inteligencia artificial en la gestión documental archivística", *Ciencia de Datos y Perspectivas de la Inteligencia Artificial*, Ramón Fernández, F. (Coord.), Tirant lo Blanch, Valencia, 2024, pp. 181 ff.

into account, which gives us a valuable perspective in relation to the quality management system and technical documentation. The sandbox, as indicated in the aforementioned legal text, enables cooperation between Artificial Intelligence users and providers, validating from both aspects the implementation of the requirements of both high-risk Artificial Intelligence systems and general purpose systems and foundational models in relation to compliance with EU requirements.

In the case of Article 11 on requirements for high-risk AI systems, which focuses on technical documentation, it is intended that the technical documentation should be available before the high-risk AI system is placed on the market or put into service. The idea is to ensure maximum safety of the system and compliance of the system with the requirements that are demanded, and highlights the reference to SMEs and start-ups, with the above-mentioned regulations applying at EU level.

One of the amendments of interest is that which is incorporated in relation to SMEs with regard to technical documentation, which can make use of the simplified form to be established by the Commission to facilitate this work for small and micro-enterprises.

Article 17 of the final text focuses on the quality management system and the adoption thereof by AI system providers, which as indicated in Royal Decree 817/2023 is any private legal person, public sector entity in Spain, or other body, which has developed or for whom an Artificial Intelligence system has been developed, and which introduces it on the market or puts it into service under its own name or trademark, whether for a fee or free of charge. The AI provider will be designated in the following ways depending on the stage of the process.

Of particular note is the reference to financial institutions and the indication that providers shall in any case respect the degree of rigour and the level of protection required to ensure the compliance of their AI systems with the Regulation.

In addition, the recent legislative resolution of the European Parliament of 9 November 2023, to which we have referred in this study, should be taken into account in relation to access to and use of data.

Article 18 of the final text is related to what is indicated in articles 11 and 17 of the same legal text and refers to the preservation of the documentation for the period of time established by the regulation. This documentation keeping of any product is a further step towards traceability and safety for the consumer.

The amendments to Article 18 of the proposal for a Regulation delete paragraphs 1 and 2 of this provision.

Article 50 of the Proposal for a Regulation, which referred to the conservation of documents, is left without content, although the precept exists with the same numbering but referring to "Transparency obligations for providers and deployers of certain AI systems", and the initial content is moved to the current Article 18 mentioned above.

THE OBLIGATION TO KEEP RECORDS OF HIGH-RISK SYSTEMS IN THE ARTIFICIAL INTELLIGENCE ACT

Wilma Arellano Toledo[1]
PhD from the Complutense University of Madrid. OdiseIA

Antonio Merchán Murillo
PhD. Lawyer. OdiseIA. Lecturer (accredited to Reader) at the University of Cadiz.[2]

I. Introduction

As is well known,[3] the European Artificial Intelligence Act classifies AI systems into several categories (*vid. supra*) and one of these is that of systems considered as high risk, which are those that inter alia "must be subject to a conformity assessment carried out by an independent body for their placing on the market or putting into service in accordance with the Union harmonisation legislative acts listed in Annex I" (formerly Annex II) of the Regulation itself and those referred to in Annex III.

It is important to note that it must be verified that all actors in the lifecycle of the high-risk AI system assume their share of the record-keeping obligations, as this may apply to developers, manufacturers, providers, importers, deployers (formerly known as users), those involved in post-marketing monitoring, etc.[4] (*see above and below*). This is because the strength of a registry or

[1] PhD from the Complutense University of Madrid. OdiseIA.

[2] PhD. Pablo de Olavide University. Lawyer. OdiseIA. Co-editor of *INTELETICA*. Assistant Professor Doctor (accredited to Full Professor) at the University of Cadiz. Member of the research team of the project "Towards a person-centred digital transition in the European Union".

[3] Article carried out with funding from the Ministry of Universities and the *NextGenerationEU* Funds of the European Union through the María Zambrano programme. Also carried out in the framework of the Project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by MCIN/AEI/10.13039/501100011033/ ERDF, EU and from the R&D&I project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) Ministry of Science and Innovation (MICINN), Proyectos de Generación de Conocimiento 2022 (oriented modality) and from the project "Towards a people-centred digital transition in the European Union", TED2021-129307A-I00, funded by MCIN/AEI/10.13039/501100011033 and by the European Union *NextGenerationEU/PRTR*.

[4] Reference can be made to the chapter dealing with Article 3 of the Regulation, concerning definitions of who each of these actors are.

log[5] lies, among other things, in the possibility to prove that they have not been altered by unauthorised parties and thus become evidence. The objective is to protect the integrity of the information, as this is highly relevant for high-risk systems to comply with their record-keeping obligations.

In the following, we will address the recitals and articles concerning the registration obligations of AI systems that will be discussed here.

II. Some preliminary notions related to record-keeping obligations

The AIA contains a large number of recitals, many of which contain elements that may be linked to the obligation of high-risk systems. The first of these is Recital 46 (formerly Recital 27), which explains that the AI systems in question can only be marketed or put into service if they comply with certain requirements, including registration. The aim is to protect two spheres: public and private. The former is designed to avoid "unacceptable risks to important public interests of the EU, recognised and protected by EU law".

In the latest version, dated May 2024, Recital 47 made specific reference to the "adverse impact" that the safety components of AI systems may have on human health and safety.

The private dimension is now set out in Recital 66, where it is provided that the obligations for high-risk AI systems (in particular record-keeping obligations) have to be fulfilled because such systems may have a detrimental effect on health, safety and fundamental rights. Similarly, the protection of the private sphere from the use of high-risk AI systems appears in Recital 46 (formerly 43) and specifies that this and other obligations also aim to avoid risks to the three aspects mentioned above (health, security and fundamental rights).

Thus, both in the protection of the public sphere and in the protection of the private dimension, registries have and will have an important role to play, to verify and even serve as means of proof and for different use cases. Certainly also for those circumstances in which a high-risk system may cause damage to a person or property, for example.

On the other hand, Recital 71 (previously in Recital 46) states that all obligations, including the obligation to generate records, must occur throughout the lifetime (previously called lifecycle) of the AI development in question.

[5] Not to be confused with the one mentioned in Recital 131 (formerly in Recital 69) concerning that high-risk system providers should be part of a register (which later lines call a database) to be managed by the Commission.

This is of vital importance, as it states that in order to meet the objectives of traceability of AI systems, "comprehensible information on how high-risk AI systems have been developed and how they perform throughout their lifetime" must be "available" and therefore "this requires keeping records and the availability of technical documentation, containing information which is necessary to assess the compliance of the AI system with the relevant requirements and facilitate post market monitoring". That is, the lifecycle of the system from its conception to the point of post-market monitoring and inspection (a chapter in this work addresses this issue in detail).

Therefore, as can be seen, an indirect reference to different actors in the processes of the whole life cycle of high-risk AI systems can be deduced, which would oblige them to adopt the imposed measures, including saving the *logs* automatically generated by these systems. By referring to the entire lifecycle, the actors can be numerous[6] and the set of *logs* can be numerous as well.

The 2021 and 2023 versions of what was Recital 46 stated that "records automatically generated by the high-risk AI system, including, for example, output data, start date and time, etc., to the extent that the system and records are under their control, should be retained for an adequate period to enable them to fulfil their obligations".

In the latest versions, those of 2024, the Recital modifies and extends the list of information to be included in the technical documentation (which also concerns the registers). However, in the version voted in March 2024

[6] And not only can the actors involved be diverse, but they can also take on different roles depending on the situations involved, since, as Recital 84 (in the previous versions in 57) states: "In order to ensure legal certainty, it is necessary to clarify that, under certain specific conditions, any distributor, importer, deployer, or other third party should be considered to be the provider of a high-risk AI system and should therefore assume all the relevant obligations. This would be the case if, for example, such a person puts his name or brand on a high-risk AI system already placed on the market or put into service, without prejudice to contractual arrangements providing for another distribution of obligations. This would also be the case if that party substantially modifies a high-risk AI system already placed on the market or put into service in such a way that the modified system remains a high-risk AI system in accordance with this Regulation, or if it modifies the intended purpose of an AI system, such as a general purpose AI system, which has already been placed on the market or put into service and which is not classified as a high-risk system, in such a way that the modified system becomes a high-risk AI system in accordance with this Regulation". Throughout the Regulation, the figures of provider, deployer, authorised representative, importer, distributor and operator (which can be the product manufacturer, deployer, authorised representative, importer or distributor, as provided for in Article 3.8 of the Regulation) appear. If we take into account that registration obligations must be adopted throughout the life cycle of the high-risk system, it can clearly be interpreted that such an obligation would be applicable to all these actors.

(retained in the May version), the Recital becomes Recital 71 and specifies that this information should "include the general characteristics, capabilities and limitations of the system, algorithms, data, training, testing and validation processes used as well as documentation on the relevant risk-management system and drawn in a clear and comprehensive form".

But, in addition, it emphasises that high-risk AI systems should "technically allow for the automatic recording of events, *by means of logs*, over the duration of the lifetime of the system" (the words in bold were added in the renumbered Recital (now 71) in the version voted in March; thus emphasising the importance of logging, but omitting the reference to the time period that appeared in earlier versions of the AIA (although this was not stipulated in specific timeframes either, but was referred to as the "appropriate" time period).

In addition, the new March and May 2024 text integrates in Recital 73 provisions that did not previously appear in this specific form, stating that "High-risk AI systems should be designed and developed in such a way that natural persons can oversee their functioning, ensure that they are used as intended and that their impacts are addressed over the system's lifecycle". It is worth mentioning that this refers to natural persons, not to users (as they were previously referred to in previous versions) nor to deployers (as they are now called), so it refers to another figure[7]. In other words, this establishes a level of transparency and human oversight that high-risk systems should comply with, that was not specified in this way in the other texts of the Regulation.

In addition, this new Recital 73 adds an extensive paragraph containing other elements involving the registers or *logs*. It mentions that the natural persons in charge of human supervision must in turn record the verifications made by each of them separately in the logs generated by the system, i.e., precisely the ones we are concerned with in this work.

Complementing the provisions of the other Recitals, Recital 82 (in previous versions it appeared in 56 or 56a) refers to cases where providers of high-risk AI systems are outside the territory of the Union, in which case they must act through the authorised representative, with the consequent technical

[7] As the recent wording of the Recital reads, it refers on the one hand to the issue of human oversight, but elsewhere it refers to persons (without the adjective "natural"), so it is understood to focus on individuals, subjects of fundamental rights (e.g., when it points out "the enormous consequences for individuals in case of incorrect matching by certain biometric identification systems"). In the concept of natural persons, the focus is on those who "have been assigned human supervision to make informed decisions about whether, when and how to intervene in order to avoid negative consequences or risks, or to stop the system if it does not work as intended".

and practical problems that this may pose for the safekeeping and "in whose hands" the records are held. The problem was, in our view, that in the 2021 version the authorised representative was given the status of "jointly and severally liable with the provider" when a product is defective (without prejudice to the applicability of the EU's own product liability rules)[8]. From a technical (and legal) point of view, this posed significant challenges as the burden placed on the authorised representative could be interpreted as excessive, as we noted in various forums.

In the 2024 version of the AIA, the authorised representative no longer appears as a jointly and severally liable person (as was the case in the previous texts until the December 2023 vote), but only as a contact person in the Community territory, for notification purposes for providers of high-risk systems based outside the European Union. However, the authorised representative will have to be appointed by written mandate and is foreseen in the current Recital 82 of the Regulation as having a primary role in ensuring the conformity of high-risk AI systems placed on the market or put into service in the Union. Thus, the obligation to generate and record-keeping does not appear here to be applicable to the authorised representative, as could be inferred from the interpretation of the previous wording of the Recital, where, being (as he was) jointly and severally liable, the record-keeping would be essential.

However, according to Article 3.5 of the latest version of the Regulation, authorised representatives are 'a natural or legal person located ["situated" in the previous version] or established in the Union who has received and accepted a written mandate from a provider of an AI system or a general-purpose AI model [of "an AI system" only in the previous version] to, respectively, perform and carry out on its behalf the obligations and procedures established by this Regulation'.

In this way, the authorised representative is once again obliged (unlike Recital 82, which does not elaborate on this) to fulfil the obligations of the high-risk AI system, which can be interpreted as clearly including the obligations to keep *logs*.

Regarding the obligations for deployers (formerly users[9]), Recital 58 (2021 and 2023 versions) stated that taking into account "the nature of AI systems

[8] Council Directive of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products. The Proposal for a Directive of the European Parliament and of the Council on liability for defective products (COM/2022/495 final) is currently under discussion.

[9] That, according to Article 3.4 of the definitions, they may be private entities, public administrations or even natural persons, provided that they "using an AI system under their own authority", except in the case of purely personal uses.

and the risks to security and fundamental rights", users must use AI systems in accordance with the instructions for use and must assume other responsibilities, including record-keeping, "as appropriate". There is a gap here, as the interpretation of this provision could also raise a number of technical and material issues, to mention only some of the implications for users, now responsible for deployment, of complying with it.

However, in the version voted in March 2024 (and in the May corrigendum) this provision appears in Recital 91, which states that deployers must comply with human supervision and record-keeping obligations and, to this end, must "r take appropriate technical and organisational measures to ensure they use high-risk AI systems in accordance with the instructions of use".

This opens up an additional window in terms of the obligations of those responsible for deployment, as they will have to implement these measures, especially the technical ones, with all that this entails in practice, since they must also ensure that the persons to whom they entrust the aforementioned obligations have the necessary skills and competences to carry them out. There is even talk of AI literacy, training and authority, which is becoming increasingly complex in relation to record-keeping.

On the other hand, Recital 133 (which had no precedent in the original version of the Regulation) refers to so-called generative Artificial Intelligence, although it does not specifically refer to it as such[10]. It states that it is desirable to define as clearly as possible when content has been created by such AI and not by a human. In order to distinguish this, and to control the legal, ethical, and technical consequences that could arise from an inappropriate and even harmful use of such generative intelligence, the providers of these systems will be required to implement a series of techniques and measures to "mark and detect" that a content comes from this type of AI and not from a human. The techniques mentioned therein include "registration methods", as appropriate and in accordance with the state of the art; this also raises a number of practical questions as to how feasible it is to actually comply with this requirement.

Finally, Recital 165, which in previous versions was Recital 81, specifies that non-high-risk AI systems could adopt codes of conduct aimed at en-

[10] It refers to the fact that "A variety of AI systems can generate large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content. The wide availability and increasing capabilities of those systems have a significant impact on the integrity and trust in the information ecosystem, raising new risks of misinformation and manipulation at scale, fraud, impersonation and consumer deception."

couraging the voluntary adoption of mandatory requirements applicable to high-risk systems in order to ensure "trusted" use of AI in the European Union. In other words, this would be a self-regulatory measure that is not binding for those systems that do not fall into the category of those we are addressing. However, this consideration, which is now moved to the aforementioned Recital 165, is qualified when it is stated that systems that are not high-risk will be encouraged to comply through codes of conduct with "all or part" of the obligations applicable to those that are, including record-keeping and governance obligations.

In addition to this, the Recital "encourages" both providers and deployers of all types of systems (high-risk and non-high-risk) to adopt the principles set out in the EU Ethical Guidelines for Trusted AI.

III. Registration obligations in the Regulation: evolution, processing and final content

In this section we will explain how the articles relating to the record-keeping obligations of high-risk systems have evolved, although it should be noted that for this particular issue of logs, the variations between the three main versions of the AIA have not been very substantial, Articles 12 (the only one that has not changed its number) and 19 (formerly 20) have received few modifications that could alter the substance of the issue and Article 26 (formerly 29) has had numerous paragraphs added, but although some of the concepts dealt with there may be linked to the obligation to keep records, the fact is that they are more oriented towards issues that are dealt with elsewhere in this work (*see above and infra*). *supra and infra*).

As we started by saying in the previous paragraph, the Artificial Intelligence Act addresses the obligations for high-risk systems to retain logs, first in Article 12 **(which appears in Section 2 - formerly Chapter 2 - of "Requirements for high-risk AI systems"),** which is the one that specifically refers to data logs[11]. Paragraph 1 states that "High-risk AI systems [previously it was written "shall be designed and developed with capabilities that enable"] shall technically allow for the automatic recording of events (logs) over the lifetime of the system [previously it was added "of the system while the high-risk AI systems are in operation"]". Although these appear to be very brief

[11] Although Article 11, concerning technical documentation, will be dealt with in depth in another chapter of this work (*see above*), it is fully related to the obligation to keep records, since such documentation must include information about the records.

nuances, the technical implications may be a little more sophisticated than the current wording at first implies, as it previously referred to the time the high-risk systems were in operation and now provides that it will be throughout the lifetime, which also implies the stage of post-marketing monitoring.

It is appropriate to explain at this point in the discussion that the registers or *logs* are files where the relevant information generated by the *kernel* or system kernel and any programme that the system may have embedded is stored. The word used is *log*[12]. As in any logbook, the log stores data on all the processes or events of the system, in this case, the high-risk AI system.

This log can contain very specific data about what happens with the system and what *kernel* outputs, such as the "timestamp" where the date and time when a certain event occurred can be recorded, but the time defined in minutes and seconds, i.e., very precisely. Therefore, the information provided by all these and many other data is highly precious and invaluable, as it certifies and proves the behaviour of the system.

In essence, "*log* files provide very important information about the implicit and explicit activities of any computer hardware and *software* system. This type of log contains all the information about the normal operation of a machine or program, helping to intercept anomalies and problems, supporting security.[13]

To make this more understandable for the non-technical reader, the *kernel* is the core of the system and is the kernel that enables "privileged mode" access to the control of a system, therefore, it is the main controller and centre of the operating system. In addition to the word kernel, the word "heart" is also used to define the kernel, and this is an indication of its importance. That is why "it is primarily responsible for mediating between user processes and the hardware available on the machine, i.e., it grants access to the hardware, to the software that requests it, in a secure manner; and parallel processing of several tasks".[14]

Now, following on from Article 12 of the AI Regulation, Article 12.2 refers to logging capabilities, which must ensure a level of traceability of the system's operation (we have already seen how this can be

[12] In all the versions of the Regulation prior to the one voted in March 2024 (which is the last one available to us), there were parts where the anglicism *logs* was written, but in the latest one the term *logs* is always used in Spanish (i.e., they are referred to as "archivos de registro" or log files) in the Spanish version of the AIA. In the English version, the term *logs* is always used.

[13] Abonyi, J. and Bántay, L. "Frequent pattern mining-based log file partition for process mining", *Engineering Applications of Artificial Intelligence*, August, No. 123, 2023.

[14] Bach, F., "Information theory with kernel methods", *IEEE Transactions on Information Theory*, 69, vol. 2, 2022 (Available at https://acortar.link/8pfUEk).

achieved, taking into account everything that is *logged* and the information it can provide), but this traceability must be possible throughout the entire lifetime[15] of the high-risk AI system. In other words, it is a level of traceability that is "appropriate to the intended purpose" (in previous versions it was "lifecycle fit for purpose").

In this way, event registration capabilities will allow (i) the detection of risk situations such as those described in Article 79 on "Procedure applicable at national level to AI systems presenting a risk" (which in turn refers to the provisions of Article 3.19 of definitions of Regulation 2019/1020[16]); (ii) facilitating post-market surveillance, in accordance with Article 72 on "Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems" (*see below*); and, (iii) that monitoring the operation of high-risk AI systems referred to in Article 26.6 below, but which essentially refers to deployers that are financial institutions.

The above discussion of Article 12 raises a number of questions, as it remains to be seen how the life cycle is consolidated in each AI system, or how it is defined and shaped. It is also open to interpretation how the phrase "appropriate to the intended purpose" is interpreted, as each actor in the life-cycle might understand it differently.

Article 12 has undergone the most relevant changes in its paragraph 3 as previous versions stated that the recording capabilities shall enable "the monitoring of the operation of the high-risk AI system with respect to: (i) the occurrence of situations that may result in the AI system presenting a risk within the meaning of Article 65(1) [now 79 above] (or lead to a substantial modification; and (ii) facilitate the post-market monitoring referred to in Article 61 [now 72 above]; (iii) the monitoring of the functioning of high-risk AI systems referred to in Article 29(4) [now 26]" (referring to human oversight,

[15] **Although this latest version does not go into detail with the term "lifetime", it does derive this from its provisions as it even mentions post-marketing follow-up and human supervision stages.**

[16] Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and product conformity and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011; Article 3.19 describes that a 'product presenting a risk' shall be considered as a product that 'may adversely affect the health and safety of persons in general, health and safety at work, consumer protection, the environment, public safety or other public interests protected by applicable Union harmonisation legislation, to an extent which goes beyond what is considered reasonable and acceptable in relation to its intended purpose or under normal or reasonably foreseeable conditions of use of the product in question, including the duration of its use and, where appropriate, its putting into service, installation and maintenance requirements".

which now appears in 26(2)). As can be seen, most of this content is now concentrated in Article 12.2.

However, now Article 12 in its third paragraph restricts the enrolment obligations only to the high-risk systems mentioned in Annex III, point 1(a) (which are remote biometric identification systems). But in addition to the restriction to such systems only, this provision also invokes a system of "minimums" when it states that remote biometric identification systems shall ensure that the enrolment capabilities shall include "at a minimum: (i) a record of the period of each use of the system[17], (ii) the reference database against whichinput data has been checked by the system, (iii) the input data for which the search has led to a match, and, (iv) the identification of the natural persons involved in the verification of the results, as referred to in Article 14.5.[18]

Article 16 (subparagraph (e), formerly (d) and in previous versions in Chapter 3, now in the so-called Section 3) on the "Obligations of providers of high-risk AI systems" provides that they shall comply with the obligation to keep "p the logs automatically generated by their high-risk AI systems as referred to in Article 19 [formerly 20]" when they are under their control. Of course, these last four words leave a lot of room for manoeuvre and a high possibility of confusion among system providers as to their record-keeping obligations.

The current Article 16 also provides in its paragraph (i) that providers of high-risk AI systems shall comply with the registration obligations referred to in Article 49(1) concerning the obligations to register in the database that the EU will manage for the purpose of monitoring these systems.

On the other hand, Article 19 (previously Article 20) expressly refers to automatically generated logs, remains with only two paragraphs and the changes have not been very noticeable, as we will see below.

In the versions that have been amended since 2019 and up to the last one in 2024, the previous Article 20 stated that logs "automatically generated by their high-risk AI systems" shall be kept. The current version (now as Article 19) provides that those "automatically generated by its high-risk AI systems" shall be kept, provided that they are under its control[19]. As can be seen, it is only a nuance in the sentence by changing the order of the words. It is also

[17] Referring to the start date and time and the end date and time of each use.

[18] In other words, it refers again to remote biometric identification systems, as Article 14 (on human surveillance - in other versions and even in other parts of the Regulation, called human monitoring) in its paragraph 5 refers again to point 1(a) of Annex III mentioned above.

[19] In previous versions, it was stipulated that records would be retained, insofar as they were under their control "by virtue of a contractual agreement with the user or otherwise by law". The latest version of this article deletes any reference to a contractual agreement.

important to note that this article refers to the fact that it is the providers of the high-risk systems that must retain these records and does not refer to those responsible for the deployment or other actors in the value chain or lifecycle of the AI system.

As regards the periods for which these log files have to be kept, in both the previous and final versions, a minimum period of six months is mentioned, 'unless provided otherwise in the applicable Union or national law, in particular in Union law on the protection of personal data'.

In other words, this Article 19 is setting a condition for retention periods that could be modified if the aforementioned log files include personal data, in which case the General Data Protection Regulation and local laws would be applicable to them, in addition to any other instrument of Union or national law that may have a bearing on such *log* retention periods.

The second paragraph of Article 19 refers specifically to providers that are financial institutions, in which case the retention of log files shall be carried out in accordance with the provisions of the legislation of that particular sector. The change that has taken place in the latest version of 2024 compared to previous drafts is that Article 74 of Directive 2013/36/EU was specifically mentioned, whereas in the final version, no reference is made to any specific legislation, but it is specified that everything shall be done in accordance with the provisions of the "relevant financial services law".

Article 26, entitled "Obligations of deployers of high-risk AI systems", formerly Article 29 "Obligations of users of high-risk AI systems", underwent many changes in the February 2024 version, compared to the previous ones, to the point of having been given an almost new wording, in some aspects, with specific changes being made later, as we will see.

In this context, paragraph 1 of this Article sets out the appropriate technical and organisational measures[20] to ensure that they use these systems in accordance with the accompanying instructions for use, in accordance with paragraphs 3 and 6, formerly paragraphs 2 and 5, in the other working versions.

These obligations should be understood, as indicated in Recital 120, as

[20] It should be noted that we are talking about measures to ensure interoperability, in a purely IT sense. This is of great importance, as we are talking about interoperability measures, in this case: a) technical: connecting systems efficiently without cybersecurity failures (e.g. data exchange) and b) organisational, referring to business processes and internal structures, e.g. that systems can exchange data beyond their technical content, i.e. we are talking about standardisation through the use of ISO standards, we are talking about *documents that contain rules, instructions or characteristics that can be used to ensure which materials, products, processes and services are fit for purpose*.

being particularly relevant to facilitate the effective implementation of Regulation (EU) 2022/2065[21], whereas they had only been included in the previous version, where they were located in Recital 70d.

As for the instructions for the use of technical and organisational measures[22], they refer, in the first place, to human supervision by natural persons having the necessary competence, training and authority, as clarified in paragraph 2 of the article, added to the latest March version. However, it should be noted that in this version a reference to "the necessary support"[23] is deleted, which relates to the "necessary competences", in particular an appropriate level of AI literacy, training and authority to properly perform such tasks, as indicated in Article 4 referring to AI literacy.

In this sense, paragraphs 3 and 4 of the Article, formerly 2 and 3 respectively, have identical wording in the previous texts in paragraphs 2 and 3, with the exception of the reference to the person responsible for deployment.

In addition, with reference to paragraph 5, formerly paragraph 4, it is introduced as indicated in Recital 91, which does not appear previously, and which has no correlation with any other recital in the previous versions, where it indicates the need to ensure the proper monitoring of the operation of an AI system in a real environment. In this context, it identifies the need to define the specific responsibilities of those responsible for the deployment. In particular, users shall monitor the operation of the high-risk AI system on the basis of the instructions of use. In addition, further obligations need to be defined in relation to the monitoring of the operation of the AI systems and record keeping, as appropriate.

In this regard, paragraph 5 indicates that deployers hall monitor the operation of the high-risk AI system on the basis of the instructions for use and, where relevant, inform providers in accordance with Article 72 (referring to post-market surveillance). Where deployers have reason to consider that the use of the high-risk AI system in accordance with the instructions may result in that AI system presenting a risk within the meaning of Article 79(1), they shall, without undue delay, inform the provider or distributor and the relevant

[21] Regulation (EU) 2022/2065 on a single market for digital services and amending Directive 2000/31/EC (Digital Services Regulation).

[22] With regard to these organisational measures, it should be borne in mind that these measures will also establish skills that are covered within the various organisations, i.e. which actors are involved and will therefore be liable, and that a "personal scope of application" will be determined, in a computer-legal sense, in accordance with the rules of liability that may be established in other instruments.

[23] This omission can be seen when comparing the text of the draft agreement issued by the European Commission, which we date back to February 2024.

market surveillance authority, and shall suspend the use of that system, we discussed the procedure applicable at national level to AI systems presenting a risk.

Similarly, if the deploying officer detects a serious incident, he shall immediately report the incident first to the provider and then to the importer or distributor and the relevant market surveillance authority. In the event that the deployer is unable to contact the provider, Article 73 (referring to the reporting of serious incidents) shall apply *mutatis mutandis*. This obligation shall not cover sensitive operational data of deployers of AI systems which are law enforcement authorities.

In the case of deployers that are financial institutions, because, as indicated in Recital 91, further obligations need to be defined in relation to the oversight of the functioning of AI systems and record keeping, as appropriate, as they are also subject to requirements relating to their governance, systems or internal processes under Union financial services legislation, the oversight obligation under the first subparagraph shall be deemed to be fulfilled when the rules on governance, systems, internal processes and mechanisms are respected in accordance with the relevant financial services law.

Paragraph 6, formerly included in paragraph 4, refers to the retention of records, as indicated in Recital 91 and Article 19 (*see above*), as appropriate. In this regard, to the extent that such records are under their control[24] for an appropriate period of time, of at least six months, unless provided otherwise in applicable Union or national law, in particular in Union law on the protection of personal data. With regard to financial institutions, logs as part of the documentation kept under Union financial services law.

Paragraphs 7 and 8, formerly paragraphs 5(a) and (b)-(c) respectively, as compared with the previous version and without being equivalent to any other version, relate to the keeping of records in paragraph 6, referring to deployers who are employers shall inform workers' representatives and the affected workers that they will be subject to the use of the high-risk AI system and deployers of high-risk AI systems who are public authorities, or Union institutions, bodies, offices or agencies shall comply with the registration obligations referred to in Article 49 (on Registration).

[24] On this point, in the text produced in the version mandated by the European Parliament, prior to the agreement, a reference was removed from the text which may be particularly interesting in that it referred to them being under its control "and are necessary to ensure and demonstrate compliance with this Regulation, for ex-post audits of any reasonably foreseeable malfunction, incident or misuse of the system, or to ensure and monitor the proper functioning of the system throughout its life cycle".

Paragraph 9, formerly paragraph 6 and unchanged from all previous versions, refers to the information provided under Article 13 to comply with their obligation to carry out a data protection impact assessment under Article 35 of Regulation (EU) 2016/679[25] or Article 27 of Directive (EU) 2016/680[26]. Paragraph 10, formerly paragraph 6a-a), is unchanged from the previous and preceding paragraphs.

Paragraph 11, formerly paragraph 6b-b), makes a change in relation to the reference Article with respect to the provisions of Article 50[27], formerly Article 52, those responsible for the deployment of high-risk AI systems referred to in Annex III who take decisions related to natural persons shall inform the natural persons that they are subject to the use of the high-risk AI systems. Furthermore, while the immediately preceding version referred to high-risk AI systems used for law enforcement purposes, it elaborates on systems that are used for law enforcement purposes or assist in making decisions relating to natural persons shall inform law enforcement, thereby implementing Article 13 of Directive (EU) 2016/680.

Finally, paragraph 12, paragraph 6quarter, c, is unchanged from the previous version, with no precedent in previous versions.

IV. Final reflections and analysis

After analysing the entire regulatory evolution of the recitals and articles concerning the obligation to retain records automatically generated by high-risk AI systems in the European Artificial Intelligence Regulation, it can be concluded that the debate among EU authorities has been intense and has led to a series of profound modifications in the different versions, but espe-

[25] General Data Protection Regulation, cited above.

[26] Article 27 of the Directive on the protection of individuals with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data refers to when a type of processing is likely, by its nature, scope, context or purposes, to involve a high risk to the rights and freedoms of natural persons, in particular using new technologies, by its nature, scope, context or purposes, is likely to result in a high risk to the rights and freedoms of natural persons, 'Member States shall provide that the controller carries out prior assessment of the impact of the envisaged processing operations on the protection of personal data'. In this regard, the reference that "controllers may rely in part on such data protection impact assessments to fulfil some of the obligations laid down in this Article, to the extent that the data protection impact assessments meet those obligations", in the text made in the version mandated by the European Parliament, was deleted.

[27] Concerning transparency obligations for providers and users of certain AI systems.

cially between the first and the last one (original of 2021 and the first vote of March 2024 and the corrigendum of May of the same year).

With regard to the record-keeping obligations explained so far, one may wonder whether all these record-keeping obligations can lead to problems, for example, when personal data are involved (we have already seen in which cases the relevant legislation is fully applicable). And we mean not only personal data and even sensitive or special categories of data (such as neurodata[28]), but only personal data.

Logs may contain confidential information that should not be disclosed for privacy reasons or even because disclosure makes the security_of the system vulnerable. In these cases it is necessary to protect the confidentiality_of the information. To solve these problems, encryption_techniques are used in the *logs*.

The obligations of those responsible for the deployment of high-risk AI systems indicate issues of relevance, as evidenced by the numerous changes that have taken place since they were first drafted. The aim is to ensure that the use of AI is carried out in a transparent, responsible and ethical manner. In this sense, it can be seen how, with the aim of ensuring compliance with the standard, it seeks to reduce potential risks and guarantee the responsible use of AI by establishing oversight and accountability mechanisms for the establishment of safe and fair AI systems.

[28] See, Arellano Toledo, W., "Los neuroderechos y su regulación" *Inteligencia Artificial. Iberoamerican Journal of Artificial Intelligence*, vol. 27, no. 73 (2024).

TRANSPARENCY AND PROVISION OF INFORMATION TO DEPLOYERS IN ARTICLE 13 OF THE ARTIFICIAL INTELLIGENCE ACT

María Estrella Gutiérrez David
Lecturer in Constitutional Law Complutense University of Madrid

I. Introduction: a general approach to Article 13 of the Regulation

While "transparency" has been one of the most frequently mentioned principles in legal doctrine, *soft law* and nascent national sectoral legislation on AI, the interpretation of its content and scope varies as to what should be transparent (e.g., the use of data, the source code, the interaction between human and AI, automated decisions, the purpose of the use of data or the application of the AI system), who are the subjects bound by the AI system and who are the stakeholders to whom transparency is addressed and, where appropriate, the purpose of transparency (e.g., harm minimisation, improvement of the quality of the data, legal reasons, confidence building, principle of democracy).[1] In this regard, there is a broad consensus that the level or degree of transparency, qualitative and quantitative, may vary depending on the stakeholders (AI system deployers, the general public, individuals or groups affected by system decisions, incident analysts, regulators, certification authorities and auditors, legal operators in the administrative or judicial field).[2]

Article 8.1 Artificial Intelligence Act (hereinafter "AIA") provides that high-risk AI systems shall comply with a number of requirements "taking into account their intended purpose as well as the generally acknowledged state of the art on AI and AI-related technologies". These requirements are provided for in Section 2 of Chapter 3 and include areas related to risk management (Article 9), data quality and governance (Article 10), technical documentation

[1] Schneeberger, D. *et al.*, "The Tower of Babel in Explainable Artificial Intelligence (XAI)", in Holzinger, A., Kieseberg, P., Cabitza, F., Campagner, A., Tjoa, A M., Weippl, E. (eds.), *Machine Learning and Knowledge Extraction. CD-MAKE 2023. Lecture Notes in Computer Science*, Springer, Cham, vol. 14065 (2023), p. 67. https://doi.org/10.1007/978-3-031-40837-3_5

[2] IEEE Standards Association, "IEEE Standard for Transparency of Autonomous Systems", in *IEEE Std 7001-2021*, pp. 18-30, 4 March 2022, doi: 10.1109/IEEESTD.2022.9726144; Information Commissioner's Office and Alan Touring Institute, *Explaining decisions made with AI*, v. 1.0.14, 17 October 2022, p. 38. https://ico.org.uk/media/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence-1-0.pdf; Government of Canada, *Directive on Automated Decision-Making (2019).* Appendix B. Impact Assessment Levels, last updated, 25 April 2023.

(Article 11), record-keeping (Article 12), transparency and provision of information to deployers (Article 13), human oversight measures (Article 14), and accuracy, robustness and cybersecurity (Article 15). The purpose of this Chapter is to systematise and analyse the content and scope of the transparency and disclosure requirement in Article 13 AIA.[3]

1. The legislative process of shaping Article 13 and its interpretative challenges

Unlike the text finally published on 14 May 2024 by the European Parliament and the Council, the initial proposal for the Regulation presented by the European Commission[4] did not include any definition of "transparency". Among the more than 700 amendments tabled by the European Parliament to the text proposed by the Commission, an additional paragraph was included to Art. 13.1 with an explicit mention of transparency. In this amendment, transparency was identified as the adoption of technical measures to ensure the interpretability of decisions taken by high-risk systems both by the provider itself and by the user of the system (in the terminology of the final version of the AIA, the "deployer"). Moreover, this definition was connected to the right to an explanation to the individual persons affected by decisions taken by AI systems with a content similar to that of the current Article 86 of the Regulation.[5]

[3] See European Parliament and Council, REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL laying down harmonised rules in the field of Artificial Intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) No 168/2013, (EU) No 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU 2016/797 and (EU 2020/1828).No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Regulation) with publication in the Official Journal of the European Union (PE-CONS 24/24) of 14 May 2024 still pending at the time of closure. At the time of closure of this Chapter, the final text (hereinafter "PE-CONS 24/24") has not yet been published in the Official Journal of the European Union.

[4] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union (COM/2021/206 final).

[5] Vid. Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union (COM(2021)0206 - C9-0146/2021 - 2021/0106(COD)), amendment No 300. The amendment stated that: "[…] transparency shall mean that, at the time of placing the high-risk AI system on the market, all technical means available in accordance with the generally recognised state of the art shall be used to ensure that the provider and the user can

Already in the provisional agreement of 2 February 2024[6] adopted by the European co-legislators during the trialogue process, a Recital (14a) was introduced defining transparency along the lines proposed by the Ethical Guidelines of the Commission's Group of Experts, the content of which was taken up verbatim in Recital (27) of the text of the Regulation finally adopted.[7]

In contrast to the Commission proposal, the text of the Regulation presented by the Council included two new paragraphs (3)(b) (i) and (3)(e), and two new paragraphs (3)(b)(va) and (3)(ea), to the text of Article 13 as foreseen in the Commission proposal.[8]

In turn, Article 13.3(iv), added to the Commission proposal, has its origin in the interim agreement of 2 February 2024[9] which, with some slight final variations, has been taken over in the definitive text adopted in May 2024.

In addressing the content and scope of the transparency requirement under Article 13 AIA, the following considerations should be made. Firstly, in relation to AI systems, "transparency" is a polysemic concept, variable, and modulable depending on the domain (technical, ethical, soft law, legal), the context of application of the system (public or private sector), its impacts (individual or collective), the purpose of the transparency (internal or external) and the subjects to whom it is addressed.[10] Secondly, although transparency is

interpret the output information of the high-risk AI system. The user shall be able to understand and properly use the AI system by knowing, in general, how the AI system works and what data it processes, enabling him to explain the decisions taken by the AI system to the person concerned in accordance with point (c) of Article 68.

[6] Vid. PE758.862v01-00.

[7] PE-CONS 24/24.

[8] See, *Council, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Regulation) and amending certain Union legislation* (CONSIL_ST_15698_2022_INIT_6 DEC 22), 6 December 2022. The indent in paragraph (3)(b)(i) included, next to the intended purpose, the following indent: "including the specific geographical, functional or behavioural environment in which the high-risk AI system is intended to be used". The new paragraph (3)(b)(va) sought to add in the instructions for use a new category of information: "where appropriate, description of the expected output information from the system". Neither of these two Council proposals were successful, but a new subparagraph 3(e), "the necessary hardware and software resources", and a new subparagraph 3(ea), referring to log files (now subparagraph 3(f)), were incorporated into the final text of Article 13.

[9] The new paragraph introduced by the Interim Agreement of 2 February 2024 stated as information to be included in the instructions for use: "(iiia) where appropriate, the capabilities and technical characteristics of the AI system to provide relevant information explaining its performance".

[10] On the meanings of transparency in the field of AI, see Cotino Hueso, L., "Transparencia y explicabilidad de la inteligencia artificial y "compañía" (comunicación, interpretabili-

an essential requirement of high-risk systems, its absence is one of the fundamental challenges of AI. It is said, not without reason, that the concept of transparency could be even more opaque than that of AI itself.[11] Therefore, different scientific fields have tried to come up with solutions to improve the transparency of AI, while articulating different but related concepts, including explainability and interpretability. However, there is no common taxonomy either within the same scientific field (Computer Science, Data Science) or between different fields (Law and Data Science). In third place, this lack of consensus in the scientific field has been transferred to other domains, such as Ethics, soft-law or Law, generating a "Tower of Babel" effect of confusing terminologies and concepts that make it difficult to identify a common scientific basis.[12] The AIA has ended up picking up some of this conceptual and terminological confusion.

2. Objectives and methodological considerations

In the light of the above considerations, this Chapter aims to provide a systematic analysis of Article 13 AIA by identifying the different dimensions of transparency, as well as the content and scope of the provision. The structure of this Chapter deals, firstly, with a systematic description of Article 13 and the correlation of its paragraphs with other provisions of the Regulation, as well as the necessary references to standardisation in order to complete the interpretation of the more technical aspects of the provision. Secondly, it analyses how the requirements of transparency, interpretability, and explainability have been integrated into the text of article 13. This analysis includes a necessary reference to the technical conceptualisation, the pre-legislative background of the text and the interrelation between the three requirements, while incorporating an exhaustive study of the (insufficient and ambiguous) approach that the AIA incorporates into interpretability and explainability. Thirdly, the exegesis of article 13 addresses the three dimensions of "transparency" that the provision incorporates: a subjective dimension (who are the obliged subjects and recipients of transparency), a formal dimension (the how or mode of compliance with the obligation), and a substantive or mate-

dad, inteligibilidad, auditabilidad, testabilidad, comprobabilidad, simulabilidad…). Para qué, para quién y cuánta", Cotino Hueso, L., Castellanos, J. (coords.), *Transparencia y explicabilidad de la inteligencia artificial*, Tirant lo Blanch, Valencia, 2022, pp. 25-65.

[11] Kiseleva, A., Kotzinos, D., De Hert, P., "Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations", *Frontiers in Artificial Intelligence*, vol. 5 (2022), p. 1. https://doi.org/10.3389/frai.2022.879603

[12] Schneeberger, D. *et al.*, "The Tower of Babel…", p. 66.

rial dimension (what information must be communicated to the subjects to whom transparency is addressed). Finally, given the markedly technical nature of the Regulation, a specific section has been included on the role of standardisation in the development of Article 13.

In establishing an appropriate methodology for analysing and interpreting Article 13, the following aspects have been taken into account. Firstly, Article 13, like much of the text of the regulation, is clearly intended to be a technical regulation. Therefore, for a correct teleological interpretation of the provision, it will be necessary to integrate it not only with other provisions of the Regulation (with the same *technical purpose*) and some specific sections of Annex IV.[13] Secondly, to ensure a proper understanding of those terms with a clear technical meaning included in Article 13 (e.g., precision, metrics, performance, transparency, interpretation, explanation, among others), in the absence of a definition in the Regulation, it has been necessary to refer to the technical standards and standardisation documents that some international and national bodies have begun to publish. This is the case of the International Organization for Standardization ("ISO")[14], the European Telecommunications Standards Institute ("ETSI"), the Institute of Electrical and Electronics Engineers Standards Association ("IEEE"), the National Institute of Standards and Technology of the US Department of Commerce ("NIST").

For methodological purposes, the list of technical standards approved or pending approval by the different standardisation bodies taken into account in the preparation of this Chapter is included below.

[13] Although Annex IV refers to the basic content of the technical documentation provided for in Article 11 of the Regulation, it should be noted that some of the categories of information listed therein overlap with the categories of information provided for in Article 13(3). As Annex IV deals in more detail with the categories of information, its reference allows, in many cases, to integrate the meaning of the different paragraphs of Article 13.

[14] At the time of writing this Chapter, the ISO/IEC JTC 1/SC 42 Committee on Artificial Intelligence has approved some 28 technical standards on Artificial Intelligence and is in the process of developing a further 30 standards. Among the technical standards under development that have been considered for the purpose of interpreting the content and scope of Article 13 are: ISO/IEC DIS 12792 - Transparency taxonomy of AI systems and ISO/IEC CD TS 6254 - Objectives and approaches for explainability and interpretability of ML models and AI systems.

Table 1. List of referenced Technical Standards

Technical Standard [Status]
ISO/IEC DIS 12792: 2024(en). Information technology - Artificial Intelligence -Transparency taxonomy of AI systems [In process].
ISO/IEC 25059:2023(E). Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality model for AI systems [Approved: 26/06/2023].
ISO/IEC 25010:2023(en). Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Product quality model [Approved: 15/11/2023].
ISO/IEC 22989:2022. Information technology - Artificial Intelligence - Artificial Intelligence concepts and terminology [Approved: 19/07/2022]. Corrigenda: ISO/IEC 22989:2022/AWI Amd 1.
ISO/IEC TS 5723:2022(en). Trustworthiness - Vocabulary [Approved 22/07/2022].
ISO/IEC 23053:2022. Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) [Approved: 20/06/2022]. Corrigenda: ISO/IEC 23053:2022/AWI Amd 1.
ISO/IEC TR 24027:2021. Information technology - Artificial Intelligence (AI) - Bias in AI systems and AI aided decision making [Adopted: 05/11/2021].
ISO/IEC TR 24029-1:2021. Artificial Intelligence (AI) - Assessment of the robustness of neural networks [Approved: 10/03/2021].
IEEE 7001-2021. Standard for Transparency of Autonomous Systems [Approved 08/12/2021].
ETSI TR 104 032 (V1.1.1) (2024-02). Securing Artificial Intelligence (SAI); Traceability of AI Models [Adopted: 2024-02].
ETSI GR SAI 007 (V1.1.1) (2023-03). Securing Artificial Intelligence (SAI). Explainability and transparency of AI processing [Approved 07/03/2023].
ETSI GR SAI 009 V1.1.1 (2023-02) Securing Artificial Intelligence (SAI); Artificial Intelligence Computing Platform Security Framework [Adopted 16/02/2023].
ETSI GR SAI 004 V1.1.1 (2020-12). Securing Artificial Intelligence (SAI); Problem Statement. [Approved: 12/2021].
NISTIR 8312 (2021). Four Principles of Explainable Artificial Intelligence.
NISTIR 8269 (2019). A Taxonomy and Terminology of Adversarial Machine Learning.

Source: Own elaboration

The consultation and reference to the technical standards listed in Table 3 *below* has allowed to complete the meaning of the more technical aspects of the content of Article 13, as well as to introduce a comparative analysis from the perspective of standardisation in terms of the categorisation of the types and levels of transparency (Table 2) or of the categories of stakeholders and relevant information (Table 7).

In particular, ISO/IEC DIS 12792:2024(en), ISO/IEC 25059:2023, on quality model for AI systems, or ISO/IEC 22989:2022, where basic concepts on AI are included, have been particularly useful.

In the case of ISO/IEC DIS 12792:2024(en), although it is a technical standard in the pipeline at the time of closure of this Chapter, its consultation is relevant for the purposes of this Chapter because of the levels of transparency and technical information it establishes (AI system context level, AI system level, AI model level, and data set level used by the system).

Table 2. Transparency levels in ISO/IEC DIS 12792:2024(en)

Taxonomy of the level of transparency	**Categories of information**
Context level [7]	General context [7.1] Social context [7.2]; General [7.2.1]; Labour practices [7.2.2]; Consumer needs [7.2.3]. Environmental context [7.3]
System-level taxonomy [8].	General [8.1] Basic information [8.2] Organisational processes [8.3]: General [8.3.1]; Governance [8.3.2]; Management system [8.3.3]; Risk management [8.3.4]; Quality management [8.3.5]. Applicability [8.4]: General [8.4.1]; Intended purposes [8.4.2]; Capabilities [8.4.3]; Functional limitations [8.4.4]; Recommended uses [8.4.5]. Excluded uses [8.4.6]. Summary of technical characteristics [8.5]: General [8.5.1]; Expected inputs and outputs [8.5.2]; Production data [8.5.3]; Logging and storage [8.5.4]; System decomposition [8.5.5]; Application programming interface [8.5.6]; Human factors [8.5.7]; Deployment methods [8.5.8]; Configuration management [8.5.9]. Access to the internal elements [8.6]. Quality and performance [8.7]: General [8.7.1]; Verification and validation processes [8.7.2]; Runtime measurements [8.7.3]; Comparison with alternative systems [8.7.4].

Model-level taxonomy [9].	General [9.1] Basic information [9.2]. Usage [9.3]: Processing performed by the model [9.3.1]; Dependency on other models [9.3.2]; Consistency with intended purposes of the AI system [9.3.3]. Technical characteristics [9.4]: Type of technology used [9.4.1]; Attributes extracted from input data [9.4.2]; Algorithm used for processing [9.4.3]; Model building procedure [9.4.4]; Hyperparameters [9.4.5]; Input and output formats [9.4.6]; Processing hardware [9.4.7]; Computational costs [9.4.8]; Models in evolutionary systems [9.4.9]. Data used [9.5] Functional correction [9.6].
Taxonomy at the dataset level [10].	General [10.1] Basic information [10.2] Data source [10.3]. Data properties [10.4] Domain and purpose of the dataset [10.5]. Biases and limitations of the data [10.6]. Social considerations [10.7] Data preparation done [10.8]. Maintenance of the dataset [10.9].

Source: own elaboration based on ISO/IEC DIS 12792:2024(en)

It is also important to note that the European Commission has mandated the European Committee for Standardisation (CEN) and the European Committee for Electrotechnical Standardisation (CENELEC) to develop ten technical standards specifying the main obligations foreseen in the AIA for high-risk systems, including a specific technical standard for transparency.[15]

[15] European Commission, *Commission Implementing Decision of 22.5.2023 on a standardisation request to the European Committee for Standardisation and the European Committee for Electrotechnical Standardisation in support of Union policy on Artificial Intelligence*, C(2023) 3215 final. Annex I. The list of new European technical standards and European standardisation documents which

II. Systematic description of Article 13: transparency dimensions

The transparency requirement provided for in Article 13 AIA has different dimensions: subjective, formal, and substantive.[16] These transparency dimensions follow a similar approach to that of other general or sectoral legislation that establishes formal and substantive transparency obligations for certain subjects falling within the scope of the rule, and more specifically the General Data Protection Regulation ("GDPR").[17] However, with respect to high-risk systems that have an individual impact on the fundamental rights of natural persons or similar legal effects, the doctrine considers that Article 13, in particular, and the Regulation, in general, would not significantly expand the content of the information obligations of Articles 13-15 or the safeguards of Article 22 of the GDPR.[18]

the Commission has mandated to CEN and CENELEC includes the following. European standard(s) and/or European standardisation document(s) on (1) risk management systems for AI systems; (2) governance and quality of data sets used to build AI systems; (3) record keeping through event logging capabilities of AI systems; (4) transparency and information to users of AI systems; (5) human monitoring of AI systems; (6) accuracy specifications for AI systems; (7) robustness specifications for AI systems; (8) cybersecurity specifications for AI systems; (9) quality management systems for providers of Artificial Intelligence systems, including post-marketing follow-up processes; (10) conformity assessment of AI systems.

[16] Cf. Kiseleva, A. *et al*, "Transparency of AI in Healthcare…", pp. 4-5.

[17] Cf. Sovrano, F., Sapienza, S., Palmirani, M., Vitali, F., *Metrics, Explainability and the European AI Act Proposal*, *J - Multidisciplinary Scientific Journal*, vol. 5, no. 1 (2022), p. 131. https://doi.org/10.3390/j5010010 Note that the systematics followed by Article 13 of the AIA is very similar to that of Articles 12-14 of the General Data Protection Regulation ("GDPR"). While Article 12(1) of the GDPR concerns the controller's formal transparency obligation in relation to how the information provided for in Articles 13-15 should be provided to data subjects, as appropriate (in a concise, transparent, intelligible and easily accessible form, in plain and plain language, in writing or by other means), Articles 13, 14 and 15 determine the *what*, i.e. the material content of the information to be provided to data subjects where the personal data have been collected from the data subject himself or from a third party other than the data subject. Precisely, the information to be provided by the controller to data subjects includes the existence of automated decisions, including profiling, provided for in Article 22 of the GDPR, and, as a minimum, the "meaningful information about the logic applied, as well as the significance and the expected consequences of such processing for the data subject".

[18] Hacker, P., Passoth, JH., "Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond", Holzinger, A., Goebel, R., et al. (eds) *xxAI - Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science*, vol 13200 (2022). Springer, Cham, p. 361. https://doi.org/10.1007/978-3-031-04083-2_17 Contrary to the authors' approach, two nuances should be introduced regarding the relationship between the Artificial Intelligence Regulation and European data protection law. First, it should be noted that, according to Article 18(9) of the Regulation, the information provided by the high-risk system provider under Article 13 shall be used by the deploying controller to comply with the obligation to

The subjective scope of application of the transparency requirement is contained in Article 13.1 and includes the identification of the obligated parties (providers) and the parties to whom transparency is addressed (deployers). Furthermore, the manifestations of "formal transparency" refer to the manner in which the material content of the information obligations under Article 13 should be presented and are contained throughout Article 13.1 and 2. Finally, "material or substantive transparency" would comprise the categories of information to be included in the instructions for use by the provider of the high-risk AI system and described in Article 13.3.

Based on this outline, the determination of the content and scope of Article 13 requires its integration with other provisions of the AIA, mainly with the definitions contained in Article 3, the provisions of Section 2 of Chapter III of the Regulation and Annex IV. For hermeneutical purposes, Table 2 below incorporates the main correspondences between Article 13, with other provisions contained in the AIA, and a list of technical standards for the purpose of specifying the content and scope of the transparency requirement.

carry out a data protection impact assessment under Article 35 of Regulation (EU) 2016/679 or Article 27 of Directive (EU) 2016/680. Secondly, the scope of application of the right to an explanation recognised by Recital (71) of the GDPR in relation to fully automated individual decisions, including profiling, with legal or similar effects of Article 22 would indeed be extended, insofar as the right to an explanation in relation to individual decisions taken by high-risk Annex III systems, provided for in Article 86 AIA, would apply in cases where this right "is not otherwise provided for in Union law" (paragraph 3). A systematic interpretation of Article 86(3) AIA leads to the conclusion that the right to an explanation would extend to individual decisions, including profiling, which are not fully automated, where there is human supervision, provided that such processing of personal data has an impact on the health, safety or fundamental rights of individuals. Precisely those that the Regulation would be excluding from the scope of Article 22 and Recital (71). Cf. Laux, J. "Institutionalised distrust and human oversight of Artificial Intelligence: towards a democratic design of AI governance under the European Union AI Act", *AI & Society. Knowledge, culture and Communication* (2023), p. 5. https://doi.org/10.1007/s00146-023-01777-z The author argues that, unlike the GPDR's approach to fully automated systems, the AI Regulation would also apply to partially automated systems, where there is human intervention.

Table 3. Correspondence between Article 13 AIA, Articulated AIA and Standardisation

	Article 13 **Transparency and provision of information to deployers**		**Articulated AIA**	**Technical reference standards**
Subjective and formal transparency	**Subjective scope of application and level of transparency ensuring interpretability of the system and compliance with obligations**	1. High-risk AI systems shall be designed and developed in such a way as to ensure that their operatation is ***sufficiently transparent*** to enable deployers to ***interpret*** a system's output and use it appropriately. An ***appropriate type and degree of transparency*** shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Section 3.	Article 3(2) (definition of "Provider"). Article 3(4) (definition of "Deployer").	Concept of "transparency" (ISO/IEC 22989:2022, 3.4.14, 3.5.15, 5.15.8; ISO/IEC DIS 12792:2024(en), 5.3; ISO/IEC TS 5723:2022(en), 3.2.19, 3.2.20; ETSI GR SAI 007 V1.1.1 (2023-03), 3, 4; IEEE Std 7001-2021, 4.1; NISTIR 8269, 3.99). Transparency levels according to stakeholder role (ISO/IEC 22989:2023, 5.19, ISO/IEC DIS 12792:2024(en), 6.2, IEEE Std 7001-2021, 5). Concept of "interpretability" [ISO/IEC CD TS 6254, under development].
Formal transparency	**Instructions for use and characteristics of the information provided**	2. High-risk AI systems shall be accompanied by instructions for use in an ***appropriate digital format or otherwise*** that include ***concise, complete, correct and clear information that is relevant, accessible and comprehensible*** to deployers.	Article 3(15) (definition of "instructions for use") Article 11 (technical documentation)	Presentation and adequacy of information and examples of transparency in AI systems [ISO/IEC DIS 12792:2024(en), 5.3, Annex A; ETSI GR SAI 007 V1.1.1 (2023-03), 4, 5, 6].
Material transparency	**Contents of the instructions for use**	3. The instructions for use shall contain ***at least*** the following information:		
	Information on the system provider	(a) the ***identity and the contact details of the provider*** and, where applicable, of its authorised representative;	Article 3.3, 3.4., 3.3 (concepts of provider, deployer, authorised representative).	Definition of stakeholder categories or roles [ISO/IEC 22989:2023, 5.19].

	Functional suitability	(b) the characteristics, capabilities and limitations of ***performance*** of the high-risk AI system, including:	Article 3(18) (definition of "operation of an AI system)".	Concept of "functional suitability" (in software and in AI systems) (ISO/IEC 25010:2023(en), 3.1; ISO/IEC 25059:2023(en), 5.1). Taxonomies of transparency levels (ISO/IEC DIS 12792:2024(en): context-level, system-level, model-level, data-level).
	Purpose	(i) its intended ***purpose***;	Article 3(12) (definition of "intended purpose").	System-level taxonomy (applicability) (ISO/IEC DIS 12792:2024(en), 8.4). Model-level taxonomy [[ISO/IEC DIS 12792:2024(en), 9.3.3].
	Predictive performance (and its metrics), robustness and cybersecurity	(ii) the ***level of accuracy*** (including the **parameters** for assessing it), robustness and cybersecurity referred to in Article 15 against which the high-risk AI system has been tested and validated and which can be expected, as well as any known and foreseeable circumstances that may affect the expected level of accuracy, robustness and cybersecurity;	Article 15 (accuracy, robustness and cybersecurity).	Concept of "functional correctness" in AI systems (ISO/IEC 25059:2023(EN), 3.2.3, 5.4), ISO/IEC DIS 12792:2024(en), 9.6). ML performance metrics (I SO/IEC 23053:2022(en), 6.5.5.] Concept of "robustness" in AI systems (ISO/IEC 25059:2023(en), 3.2.5, 5.5); in neural networks and methods for their measurement (ISO/IEC TR 24029-1:2021(en), 3.6, 4.1.1, 5, 6, 7). Securing Artificial Intelligence (SAI): problem statement [ETSI GR SAI 004 V1.1.1 (2020-12)]; AI computing platforms [ETSI GR SAI 009 V1.1.1 (2023-02)].

	Circumstances giving rise to risks to health, safety or fundamental rights	(iii) any ***known or foreseeable circumstance,*** related to the use of the high-risk AI system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to ***risks to the health and safety or fundamental rights*** as referred to in Article 9(2);	Article 3(13) (definition of "reasonably foreseeable misuse").	Taxonomy of the "context-level" (ISO/IEC DIS 12792:2024(EN), 7) and of the data set level (SO/IEC DIS 12792:2024(en), 10.6, 10.7).
	Explainability	(iv) where applicable, the ***technical capabilities and characteristics*** of the high-risk AI system to provide ***information that is relevant*** to explain its output;	Annex IV 2(a), concerning methods and measures. Annex IV 2(b), regarding logic, assumptions, parameters, trade-offs.	Explainability and transparency: ETSI GR SAI 007 V1.1.1 (2023-03).
	Functional suitability in relation to persons/groups concerned	(v) when appropriate, its ***performance regarding specific persons or groups of persons*** on which the system is intended to be used;	Annex IV. 3	Context-level taxonomy (ISO/IEC DIS 12792:2024(en), 7).
	Specifications for input data and training, validation and test datasets	(vi) when appropriate, specifications for the ***input data***, or any other relevant information in terms of the ***training, validation and testing data sets used***, taking into account the intended purpose of the high-risk AI system;	Article 3(29) (definition of "Training data"). Article 3(30) (definition of "Validation data"). Article 3(31) (definition of "Validation data set"); Article 3(32) (definition of "Testing data"). Article 10 (data and data governance) Annex IV. 2(d), in relation to input data. Annex IV. 2(g), including validation and test data.	Taxonomy of the "data-set-level" (ISO/IEC DIS 12792:2024(en), 10). Biases in AI systems [ISO/IEC/TR 24027:202 1].
	Interpretability	(vii) where applicable, ***information to enable deployers to interpret*** the output of the high-risk AI system and use it appropriately.	Annex IV. 2	Transparency and explainability (EEA Std 7001-2021, 3.1, 4.1; ETSI GR SAI 007 V1.1.1 (2023-03); NISTIR 8312 (2021)).

	Modifications and functional suitability p or defect	(c) the ***changes to the high-risk AI system and its performance which have been pre-determined*** by the provider at the moment of the initial conformity assessment, if any;	Article 3(23) (definition of "Substantial Change"). Annex IV. 2(f). Annex IV. 6. Annex IV. 8.	ISO/IEC DIS 12792:2024(en), 8.4, 8.7, 9.4.9.
	Human surveillance measures to facilitate interpretability/ explainability	(d) the ***human oversight measures*** referred to in Article 14, including the ***technical measures put in place to facilitate the interpretation*** of the outputs of the high-ri,sk AI systems by the deployers;	Article 14 (human surveillance). Annex IV.3, regarding technical measures put in place to facilitate the interpretation of output information. Annex IV. 2(e) Annex IV. 3.	ISO/IEC DIS 12792:2024(en), 8.5.7.
	Performance efficiency	(e) the computational and hardware resources needed, the expected lifetime of the high-risk AI system and any necessary maintenance and care measures, including their frequency, to ensure the proper functioning of that AI system, including as regards software updates;	Annex IV. 1(c), in respect of updates. Annex IV. 2(c), regarding computational resources.	ISO/IEC DIS 12792:2024(en), 9.4.7, 9.4.8, 9.4.9.
	Traceability	(f) where relevant, a description of the mechanisms included within the high-risk AI system that allows deployers to properly collect, store and interpret the logs in accordance with Article 12.	Article 12 (Record-keeeping). Annex IV 2(g) (test logs).	Event logging [ISO/ IEC DIS 12792:2024(en), 8.5.4].

Source: Own elaboration.

As can be seen in Table 3, given the highly technical nature of Article 13, and pending the development by CEN and CENELEC of the corresponding technical standard in application of the European Commission's implementing decision of 22 May 2023 (C(2023) 3215 final), the delimitation of the content and scope of Article 13 has been completed on the basis of other technical standards published by other standardisation bodies relating to AI systems.[19]

[19] In particular, the following technical standards have been referenced: ISO/IEC DIS

The doctrine has been critical of the increasingly widespread and debatable trend of the European Commission to delegate the process of specifying legal norms to private law bodies (subscription and intellectual property retention model, greater exposure to lobbying, lack of democratic control, harm to European consumers and small developers in accessing the norms).[20]

In any case, the possible interpretations of Article 13 of the Regulation included in this Chapter could be expanded, qualified and even corrected by the content of the technical standard on "transparency and provision of information" to be developed by CEN and CENELEC.

III. Transparency, interpretability and explainability: their (a) systematic treatment in Article 13

Although there have been attempts in the scientific-technical field to establish conceptual independence between the concepts of "transparency", "interpretability" and "explainability", there is no general consensus on their

12792:2024(en) (AI system transparency taxonomy); ISO/IEC 25059:2023(E) (AI system quality and assessment); ISO/IEC 25010:2023(en) (system and software quality and assessment); ISO/IEC 22989:2022 (AI concepts and terminology); ISO/IEC TS 5723:2022(en) (transparency concept); ISO/IEC 23053:2022(E) (ML-based systems framework); ISO/IEC/TR 24027:2021 (biases); ISO/IEC TR 24029-1:2021(E) (robustness of AI systems); ETSI GR SAI 007 V1.1.1 (2023-03) (explainability and transparency); ETSI GR SAI 009 V1.1.1 (2023-02) (safety in AI systems); ETSI GR SAI 004 V1.1.1 (2020-12) (safety in AI computing platforms); IEEE Std 7001-2021 (transparency); NISTIR 8312 (2021) (explainability) NISTIR 8269 (2019) (ML concepts).

[20] The doctrine has pointed out that it will be in standardisation that the real elaboration of rules that give concrete form to the application of AIA will take place. However, there is criticism of the European Commission's increasingly widespread and debatable tendency to delegate the process of specifying legal standards to private law bodies (model of subscription and retention of intellectual property, greater exposure to lobbying, lack of democratic control, prejudice to European consumers and small developers in terms of access to standards). See, among others, Schneeberger, R. Röttger, F. Cabitza *et al.*, "The Tower of Babel…", Op. cit., 75; Smuha, N. A., Ahmed-Rengers, Emma et al., *How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act*, 5 August 2021, p. 54. http://dx.doi.org/10.2139/ssrn.3899991; Veale, M., Zuiderveen, B. F., "Demystifying the Draft EU Artificial Intelligence Act - Analysing the good, the bad, and the unclear elements of the proposed approach", *Computer Law Review International*, vol. 22 (2021), pp. 97-112. http://dx.doi.org/10.9785/cri-2021-220402 The authors criticise the European Commission's increasingly widespread and debatable tendency to delegate the process of fleshing out legal rules to private law bodies, to the clear detriment of European consumers. In fact, they point out that it will be in standardisation that the real elaboration of rules for the implementation of the AIA will take place.

meaning.[21] Thus, for example, transparency often overlaps with other technical properties such as reproducibility[22], traceability, verifiability, usability, explainability and interpretability, accountability, quality or reliability of the system.[23]

Especially in the XAI domain, the concepts of "transparency" and "interpretability"[24] or "interpretability" and "explicability"[25] are often used interchangeably. Some also consider "explainability" as an integral part of "transparency"[26]. In other cases, a conceptual differentiation is sought by identifying the correlations between them.[27] Finally, part of the scientific lit-

[21] UK Parliament POST, "Interpretable machine learning", *Postnote*, no. 633, The Parliamentary Office of Science and Technology, Westminster, London, October 2020. https://post.parliament.uk/research-briefings/post-pn-0633/

[22] The concepts of "reproducibility" and "replicability" must be differentiated. In the area of machine learning, the training process is "reproducible", if under the same training setup (e.g. same training data set, code, environment), the trained model produces the same results under the same evaluation criteria. Evaluation criteria can be defined for a sample of data (e.g., inference results) or over a distribution of data (e.g., performance metrics). Reproducibility of the model training process for error elimination, model evaluation and traceability, as well as auditing and verification of claims. Reproducibility' should be distinguished from 'replicability', which means that under a different data sample (with the same distribution as the original data sample) combined with the original code and analysis, similar results are obtained. See, ETSI TR 104 032 V1.1.1 (2024-02), p. 26.

[23] Cf. ISO/IEC DIS 12792:2024(en). Information technology -Artificial Intelligence-Transparency taxonomy of AI systems [currently in process], 5.3.

[24] Barredo Arrieta, A., Díaz-Rodríguez, N., *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", in "Information Fusion", vol. 58, 2020, p. 84. https://doi.org/10.1016/j.inffus.2019.12.012

[25] See, HLEG, *The assessment list for trustworthy Artificial Intelligence (ALTAI) for self assessment*, European Commission, 17 July 2020, p. 27; Molnar, Ch., *Interpretable machine learning: A guide for making black box models explainable*, 2nd ed, 2020. https://christophm.github.io/interpretable-ml-book/; Carvalho, D. V.; Pereira, E. M.; Cardoso, J. S. "Machine Learning Interpretability: A Survey on Methods and Metrics", *Electronics*, vol. 8, no. 8, 832 (2019), p. 7. https://doi.org/10.3390/electronics8080832; Adadi, Amina and Berrada, Mohammed, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)", *IEEE Access*, vol. 6 (2018), pp. 52141-52142. DOI: 10.1109/ACCESS.2018.2870052.

[26] Winfield, Alan F. T., Booth, Serena, Dennis, Louise A., et al. "IEEE P7001: A Proposed Standard on Transparency", *Frontiers in Robotics and AI*, vol. 8, 2021, p. 3. DOI: 10.3389/frobt.2021.665729.

[27] Cf. Doshi-Velez, Finale and Kim, Been, *Towards A Rigorous Science of Interpretable Machine Learning*, 2 March 2017, p. 1. https://arxiv.org/abs/1702.08608; Lepri, Bruno, Oliver, Nuria, Letouzé, Emmanuel, *et* al. "Fair, Transparent, and Accountable Algorithmic Decision-making Processes", *Philosophy & Technology*, vol. 31, 2018, pp. 619-622. https://doi.org/10.1007/s13347-017-0279-x; Rudin, C., *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, p. 2, 2019. https://arxiv.org/abs/1811.10154; Mit-

erature identifies as a primary objective the development of AI models that are interpretable (understandable or intelligible to a human observer) from the information provided by the model itself (transparency)[28] or from complementary techniques (usually other algorithmic models) that allow explanations to be extracted (explainability) in those cases in which the model is not interpretable by a human (black boxes)[29]. In fact, this is the interpretation adopted in this chapter.

From the point of view of *soft law*, the work of conceptual distinction has not fared much better either. At the European level, for example, the terminological confusion described above has been evident in some of the documents produced by the European Commission's High Level Expert Group (hereinafter "AI HLEG"), such as the "Ethical Guidelines for Trustworthy AI" (hereinafter "Ethical Guidelines")[30] or the Self-Assessment Checklist

telstadt, B., Russell, C. and Wachter, S., "Explaining explanations in AI", *Proceedings of Fairness, Accountability, and Transparency (FAT*)*, ACM Digital Library, 2019, p. 280. https://doi.org/10.1145/3287560.3287574; Longo, L.; Brcic, M.; Cabitza, F. *et al.* "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions", *Information Fusion*, vol. 106 (2024). https://doi.org/10.1016/j.inffus.2024.102301

[28] Rudin, C., Chen, C., Chen, Z., *at al.* "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges", *Statistic Surveys*, vol. 16 (2022), pp. 1-16. DOI: 10.1214/21-SS133.

[29] Information Commissioner's Office, Alan Turing Institute, *Explaining decisions*…, *Op. cit.*, p. 69.

[30] Cf. Hleg, *Ethical Guidelines for Reliable AI*, Brussels, European Commission, 8 April 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai It should be noted that the Ethical Guidelines do not include a clear definition of these concepts, nor do they establish a proper correlation between them. Thus, for example, the Guidelines take a quasi-tautological approach to the concepts of "explainability" and "transparency". Thus, while the ethical principle of explicability would include, among other aspects, the transparency of processes and the ability to explain decisions to affected parties, the requirement of transparency would in turn include explainability. It is not clear, in any case, the meaning and scope of the different terminology used in the English version, "explicability" (ethical principle) and "explainability" (element of transparency), which in the Spanish version are translated in the same way. In fact, the terms "explicability" and "explainability" do not appear in the various English reference dictionaries (Cambridge, Oxford, Merriam-Webster), although the terms "explication" and "explanation" do, with their respective meanings including subtle differences. For its part, the concept of "interpretability" is practically absent in the principles and requirements of the Guidelines (neither linked to the ethical principle of explainability nor to the transparency requirement), although a principle of interpretability by design (from the conception of the system) and by default (adoption of the simplest and most interpretable models possible) is introduced, which is associated with the checklist of explainability (as an element of the transparency requirement).

(“ALTAI”)[31]. However, all of this lack of definition and terminological confusion has ended up largely reflected in the AIA[32], which has incorporated the Ethical Guidelines as part of its interpretative acquis.[33]

1. Meaning and types of transparency in the regulation and in Article 13

From a technical point of view, there is no uniform concept of “transparency” of AI models. In the field of XAI, transparency has been identified with the degree of intrinsic interpretability of an AI model.[34] In the field of standardisation, the concept refers either to the communication or availability to interested parties of relevant information about the system[35], or to the degree of openness of a system so that it can be inspected and not have hidden parts.[36]

Furthermore, ISO/IEC DIS 12792:2024(en) distinguishes between transparency as a technical property of the AI system (“system transparency”) and transparency as an organisational property (“organisational transparency”). With regard to transparency of the AI system, this means that relevant information about the system is made available to interested parties. Such relevant information may include: the objectives of the system, its known limitations,

[31] HLEG, *The assessment list for trustworthy Artificial Intelligence (ALTAI) for self-assessment*, European Commission, 17 July 2020, pp. 26-27. The Expert Group document does not define “transparency”, but it does define “explainability” and “interpretability”. “Explainability” would then be that “property” or “characteristic of an AI system that is intelligible to non-experts”, so that an AI system is intelligible “if its functionality and operations can be explained in a non-technical way to a non-expert.” The concept of “interpretability”, on the other hand, seems to be assimilated to that of “explainability”, by being identified with “comprehensibility, explainability or understandability”, so that “when an element of an AI system is interpretable, it means that it is possible, at least for an external observer, to understand it and find its meaning”.

[32] Kiseleva, A. *et al*, “Transparency of AI in Healthcare… “, pp. 2-3.

[33] Cf. Recitals (7), (27) and (165) of the AIA.

[34] An AI model is considered transparent if the overall performance of the model (‘simulatability’), its individual components (‘decomposability’) and its learning algorithm (‘algorithmic transparency’) are intelligible or understandable to a human. Therefore, the overall transparency of a model will depend on an appropriate balance between these three levels. Cf. Lipton, Z. C. “The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery”. *ACM Queue*, vol. 16, no. 3 (2018), p. 12; Lepri *et al.*, “Fair, transparent…”, *op. cit.*, p. 619; Barredo *et al.*, *Explainable Artificial Intelligence (XAI)…* op. cit, pp. 88-100; Information Commissioner’s Office and Alan Turing Institute, *Explaining decisions…*, pp. 61-63, 115-118; Mittelstadt, B.; Russell, Chris; Wachter, S., “Explaining Explanations…”, *Op. cit.*, p. 280.

[35] ISO/IEC 22989:2022, 3.5.15, 5.15.8.

[36] ETSI GR SAI 007 V1.1.1 (2023-03), p. 6.

design choices and prior assumptions, features, models, algorithms, training methods, details of the data used, and quality assurance processes. However, transparency needs may be different for different stakeholders. With regard to organisational transparency, it relates to how activities and decisions are communicated to relevant stakeholders and its relevance is determined because the underlying organisational principles and processes affect the AI system throughout its lifetime.[37]

In connection with this concept of "system transparency", it is important to note that the recent ISO/IEC 25059:2023(en) standard on the quality model of AI systems differentiates transparent from opaque systems depending on how the systems document, record or display information about their processes technically.[38]

In contrast to interpretability and explainability, the Regulation adopts a definition of "transparency" in its Recital (27) which is based on the Ethical Guidelines of the Commission's Expert Group.[39]

> "According to the AI HLEG guidelines […] [t]ransparency means that AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights."

In reality, this approach does not contain a definition of "transparency" as such, but rather an identification of its constituent elements, namely traceability, explainability and communication of relevant information to deployers and to persons exposed to AI systems. In addition to a lack of a definition of "transparency" *per se*, two important limitations in the definition incorporated in the explanatory part of the Regulation should be highlighted.

Firstly, the absence of the requirement of interpretability of high-risk systems in Recital (27) is surprising, unless, for the European legislator, trans-

[37] ISO/IEC DIS 12792:2024(en), 5.3.

[38] ISO/IEC 25059:2023(en), 5.6. According to the standard, transparent AI systems document, record or display their internal processes using introspection tools and data archives. The data flow can be traceable at each step, with decisions applied, exceptions and rules documented. Sequential processes can be tracked and recorded as data varies, as well as errors. Highly transparent AI systems can be built from well-documented subcomponents whose interfaces are explicitly described, ultimately making it easier to investigate system failures. In contrast, a non-transparent system has internal workings that are difficult to inspect externally. The unavailability of detailed processing records can impair the verification and assessment of social and ethical impact and the treatment of risks.

[39] Cf. Ethical Guidelines, paragraphs 75-78.

parency and interpretability are one and the same thing.[40] Secondly, the identification of the constituent elements of transparency in the explanatory part is not, however, in line with the approach of Art. 13 of the Regulation, where interpretability is mentioned on several occasions, but not explainability. Either because the latter requirement is mentioned indirectly (paragraph 3.b iv), or because it is diluted or confused with that of interpretability (paragraph 3.d).

With regard to other elements of the definition of "transparency"− traceability and communication, included in Recital (27), it should be noted that the traceability[41] of AI systems is manifested through various documentation and logging obligations established throughout the Regulation[42] provided for, inter alia, in Articles 11 (technical documentation), 12 (record-keeping), 13.3 (instructions for use), 18 (duty of the provider to retain technical documentation for a period of 10 years).

For their part, the obligations to communicate relevant information provided for throughout the Regulation make it possible to identify two types of transparency, one internal transparency of a very technical content and scope ([1], [2], [6]), and another external transparency aimed at the general public ([3], [4], [5]). In particular, the communication of relevant information is present in, among other provisions of the Regulation, the following:

- The obligation of the provider to make available to the deployer the instructions for use with the content prescribed in Art. 13(3) [1].

- The obligation of the provider to cooperate with the competent authorities under Article 21 of the Regulation to provide those authorities with all information and documentation necessary to demonstrate the compliance of the high-risk AI system with the requirements set out in Section 2 of Chapter III, as well as access to the automatically generated logs by the AI system, to the extent that such records are under the control of the provider [2].

- The obligation of the deployer to inform natural persons who are exposed to Annex III high-risk systems contained in Article 26(11) AIA. In the particular case of high-risk AI systems used by competent authorities for the

[40] It seems obvious that this limitation is rooted in the approach taken by the Commission's Expert Group's Ethical Guidelines where interpretability is absent when it comes to delimiting the elements of transparency (traceability, explainability and communication).

[41] Vid. ETSI TR 104 032 V1.1.1 (2024-02), pp. 13-14. Traceability can be understood as the tracking of the entire life cycle of a model, which includes not only the tracking of a model, but also of its data and metadata, as well as the details of the training process. Traceability can be applied to the training and test data set, the training parameters, the model, and the underlying system architecture.

[42] Cf. Recital (71).

purpose of prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, this information obligation is to be modulated in accordance with the provisions contained in Article 13 of Directive (EU) 2016/680 [3].

- The information obligation of providers and deployers to natural persons who interact with certain AI systems (systems that interact directly with natural persons, generative AI systems, emotion recognition or biometric categorisation systems, ultra-counterfeiting systems) provided for in Article 50 [4].

- Public access to the information contained in the European Commission's database for high-risk AI systems listed in ANNEX III provided for in Article 71 [5], with the exception of the non-public secure section referred to in Articles 49(4), and Article 60(4)(c) [6].[43] Annex VIII of the Regulation lists the categories of information to be included in the database according to the classification made by Articles 49 and 71 of the Regulation.[44]

The transparency provided for in Article 13 is internal and technical transparency. The purposes of transparency set out in the provision would be determined by the roles of the provider and the deployer. In the case of the provider, the purpose of the (active) transparency obligation would be regulatory compliance; while in the case of the deployer, the (passive) transparency aims not only at regulatory compliance but also at an enabling purpose in the terms that will be explained below.

2. Interpretability in Article 13: discouraging black box models?

The concept of "interpretability" is not free from terminological confusion, as in some cases it is identified or confused with explainability. Indeed,

[43] High-risk systems in the field of law enforcement, migration management, asylum and border control are essentially exempted from public access.

[44] The database is divided into three accessible Sections (A, B, and C). For example, for Annex III systems, included in Section A of the database, with the exception of critical infrastructure systems, the provider or his authorised representative shall provide, inter alia, the following information: identification and contact details and location of the provider or, where applicable, the authorised representative; the trade name of the AI system and any additional unambiguous reference allowing its identification and traceability; the description of the intended purpose of the AI system and of the components and functions supported through it; a simple and concise description of the information used by the system (data, inputs) and its operating logic; the status of the AI system (placed on the market or put into service, no longer placed on the market or put into service, recovered); a copy of the EU declaration of conformity; the electronic instructions for use; and, optionally, a URL for further information.

some include explainability as an element of interpretability.[45] "Interpretability" is a *passive* characteristic of an AI model that refers to the degree to which the behaviour and results of a particular model are understandable or intelligible to the human observer. The interpretability of a model is higher if it is easy for a person to reason and trace in a coherent way why the model made, for example, a particular prediction. In comparative terms, given a model A, it will be more interpretable than another model B if A's predictions are easier to understand than those made by B.[46] As opposed to transparency, the opacity of a model (whether intentional or intended by the designer, due to lack of training and technical skills or intrinsic to the model itself) is known in the *machine learning* community as the "interpretability problem".[47]

In the technical domain of XAI, a distinction is thus made between models that are "interpretable by design" ("transparent models") and those that are not interpretable, *prima facie*, but can be explained (and therefore interpretable) by means of different techniques consisting of extracting relevant information from the model and generating explanations.[48] The explanations generated from the model may, in turn, be of different types (mathematical,

[45] Cf. Altai, *Op. cit.*, p. 27. "Interpretability refers to the concept of comprehensibility, *explainability* or understandability. When an element of an AI system is interpretable, it means that it is possible at least for an external observer to understand it and find its meaning [italics ours]". See also, UK Parliament POST, "Interpretable machine learning", Op. cit., p. 2, where the UK Parliament notes that the concept of "interpretability" is often used to "describe the *ability to present or explain the decision-making process* of an AI intelligence system in terms understandable to humans (including AI developers, users, purchasers, regulators or those affected by the system's decisions) [emphasis added]".

[46] Barredo *et al.*, "Explainable Artificial Intelligence (XAI)…". *Op. cit.*, p. 84; Carvalho, D. V.; Pereira, E. M.; Cardoso, J. S., "Machine Learning Interpretability: A Survey on Methods and Metrics". *Electronics*, vol. 8, no. 8: 832 (2019), p. 10; Molnar, C., *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*, Leanpub, 2019, sp.

[47] An AI model is considered transparent if the overall performance of the model ('simulatability'), its individual components ('decomposability') and its learning algorithm ('algorithmic transparency') are intelligible or understandable to a human. Therefore, the overall transparency of a model will depend on an appropriate balance between these three levels. Cf. Lipton, Z. C. "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery". *ACM Queue*, vol. 16, no. 3 (2018), p. 12; Lepri *et al.*, "Fair, transparent…", *op. cit.*, p. 619; Barredo *et al.*, *Explainable Artificial Intelligence (XAI)…* op. cit, pp. 88-100; Information Commissioner's Office and Alan Turing Institute, *Explaining decisions…*, pp. 61-63, 115-118; Mittelstadt, B.; Russell, Chris; Wachter, S., "Explaining Explanations…", *Op. cit.*, p. 280.

[48] Mittelstadt, B., Russell, C., Wachter, S. (2019). "Explaining Explanations in AI". *FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, p. 280; DEEKS (2019). "The judicial demand…" pp. 1832; Barredo *et al.*, "Explainable Artificial Intelligence (XAI)…", *Op. cit.*, p. 83.

statistical, in natural language) depending on the addressee of the explanation (authority, external auditor, users of the system, affected, general public).

The relevance of explainability is justified by a widespread acceptance of the inverse relationship between interpretability and the performance of AI models. This means that simpler models tend to be more interpretable, but have a lower predictive capability; and, conversely.[49] To solve the problem of interpretability/predictive performance of AI models and, in particular, of "black box models", the XAI has been developing a set of techniques aimed at generating more explainable models while maintaining high levels of performance.[50]

However, this approach has been criticised by some authors, insofar as it only encourages the development and implementation of proprietary black box models rather than "default interpretable models" in highly critical sectors such as criminal justice or healthcare.[51] Indeed, the Information Commis-

[49] Barredo, A.; Díaz-Rodríguez, N., *et al.* "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI". *Information Fusion*, vol. 58 (2020), p. 100; Marcinkevics, R., Vogt. J. E., "Interpretability and Explainability: A Machine Learning Zoo Mini-tour", *ArXiv abs/2012.01805* (2020), p. 2; iDANAE Chair. *Interpretability of Artificial Intelligence Models, Universidad Politécnica de Madrid*, Management Solutions, Quarterly Newsletter, 2019, p. 4; Kroll, J. A., Huey, J., Barocas, S. *et al.*, "Accountable Algorithms", *University of Pennsylvania Law Review*, vol. 165, no. 3 (2017), pp. 658-660; Edwards, L. and Veale, M., "Slave to the algorithm? Why a "right to an explanation" is probably not the remedy you are looking for", *Duke Law & Technology Review*, vol. 16, no. 18 (2017), p. 8; Ananny, M. and Crawford, K., "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability", *New Media & Society* (2016), p. 983.

[50] Guning, D., *Explainable Artificial Intelligence* (XAI), [7] D. Technical Report, Defense Advanced Research Projects Agency (DARPA), 2017.

[51] Rudin, C., "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". *Nature Machine Intelligence*, vol. 1 (2019), pp. 206-215. In relation to machine learning strategies, the author questions the widespread belief that there must necessarily be a trade-off between predictive performance and interpretability, and thus the need to implement complex black box models to obtain maximum predictive performance. According to Rudin, this would not be the case when using structured data with representative and relevant attributes, as in such cases, there is usually no significant difference in performance between, for example, more complex classifiers (deep neural networks, boosted decision trees, random forests) and much simpler classifiers (logistic regression, decision lists). The author believes that instead of implementing inherently interpretable models, there is a growing tendency to promote "explainable ML" approaches, generating a second (*post-hoc*) model to explain the first black box model. For the author, this approach is problematic for two reasons. First, because the resulting explanations in *post-hoc* models can be misleading for several reasons: (i) they are not always faithful to what the original model has computed; (ii) they do not provide sufficient detail of what the black box model is doing; (iii) they may not be compatible with situations where information external to the system needs to be combined

sioner's Office and the Alan Turing Institute recommend that organisations prioritise the use of systems based on default interpretable models where possible, especially where AI systems have a potentially high adverse impact on people or are safety critical.[52]

As noted *above*, whereas in the definition of "transparency" in Recital (27) interpretability is absent in favour of explainability, in Article 13 the approach is precisely the reverse. The Regulation's approach to interpretability in Article 13 has also been criticised, insofar as it does not clarify whether it is sufficient for the provider to technically ensure the interpretability of the high-risk system in order to comply with the transparency requirement.[53]

But even if this were true, Article 13 would also not clarify the level of interpretability required. The latter is essential, as even the most interpretable models (e.g., decision trees, linear/logistic regression, Bayesian), have different levels of transparency depending on the degree of simulability, decomposability or algorithmic transparency that each model allows.[54] A list of interpretable models and the levels of transparency in relation to the elements of interpretability is given in the table below.

Table 4. Relationship between transparency and interpretable models

Model	Simulatability	Decomposability	Algorithmic transparency
Linear/logistic regression	Predictors are readable and interactions between them are minimised.	The variables remain readable, but the number of interactions and predictors involved in them has grown to force the decomposition.	If the variables and interactions are too complex, mathematical tools are needed to analyse them.

with a risk assessment; (iv) they may lead to overly complex decisions that are prone to human error. Secondly, because they encourage the proliferation of proprietary models subject to enforceable intellectual and industrial property rights in the face of the need to "open up" the model for the exercise of rights by those affected by adverse decisions in critical contexts.

[52] Cf. Information Commissioner's Office, Alan Turing Institute, "*Explaining decisions…"*, *Op. cit.*, pp. 68-69.

[53] Vid. Kiseleva, A., "Making AI's transparency transparent: notes on the EU Proposal for the AI Act", *European Law Blog*, 20 July 2021. https://europeanlawblog.eu/2021/07/29/making-ais-transparency-transparent-notes-on-the-eu-proposal-for-the-ai-act/#:~:text=Transparency

[54] Barredo, et al., "Explainable Artificial Intelligence (XAI)…". *Op. cit.*, p. 90.

Decision trees	A human can simulate and obtain the prediction of the decision tree on their own, without the need for a mathematical basis.	The model comprises rules that do not alter the data, and preserves its readability.	Human-readable rules that explain the knowledge learned from the data and allow for a direct understanding of the prediction process.
Nearest neighbours K	The complexity of the model (number of variables, their comprehensibility and the measure of similarity) matches human capabilities for simulation.	If the number of variables is too high and/or the measured similarity is too complex to be able to simulate the model completely, the similarity and the set of variables can be decomposed and analysed separately.	If the similarity measure cannot be decomposed and/or the number of variables is high, the user must resort to mathematical and statistical tools to analyse the model.

Source: Barredo *et al.* (2020)

Interpretability is mentioned throughout Article 13: in paragraph (1), paragraph (3)(b)(vii) and paragraph (3)(d). While Article 13(1) would include a mandate to providers to design interpretable high-risk AI systems in order for deployers to "correctly *interpret* and use their output information appropriately", the other two paragraphs (3)(b)(vii) and (3)(d) would refer to the content that should be included in the instructions for use to enable the deployer to interpret the output results and use the system appropriately.

One possible interpretation of the approach taken by Article 13 is that the provision seeks to promote interpretable systems by default, rather than black box systems that need techniques and tools complementary to explainability. Moreover, if one looks at the background of the Regulation, in the Ethical Guidelines, as explained *above*, interpretability is absent when defining the requirement of transparency (traceability, explainability and communication). However, in the verification checklist for reliable AI, a good part of the verification questions included in the section on explainability are aimed at checking the degree of interpretability that, by default, the system would have, and to a lesser extent, the system's capacity to generate explanations that make it possible to understand or interpret the system's results.[55]

[55] See Ethical Guidelines, op. cit., pp. 37-38.

Table 5. Checklist for reliable AI: explainability and interpretability

Explainability	
Verification question	**Purpose**
Have you assessed the extent to which the decisions and thus the outcome produced by the AI system are *understandable*?	Default interpretability
Has it been ensured that an *explanation* can be developed that is understandable to all users who may wish to know why a system made a particular decision that led to a specific outcome?	Explainability
Did you design the AI system with interpretability in mind from the beginning?	Interpretability by design
Did you research and try to use the simplest and most interpretable model possible for the application in question?	
Have you assessed whether you can analyse training and test data, and can you change and update it over time?	
Have you evaluated whether, after training and development of the model, or if you have access to the internal workflow of the model?	

Source: Own elaboration based on HLEG Ethical Guidelines (2018).

In view of this background and the fact that Article 13 itself seems to require the provider to design AI systems that enable deployers to interpret the output information and make appropriate use of the system, it would be reasonable to ask whether the European legislator's intention was not to promote *interpretable systems by default* as opposed to black box models that require additional explanation techniques.

Some authors reject such an interpretation, since restricting complex black box models in favour of interpretable models could limit innovation.[56] This argument is not without reason, as references to innovation are constant throughout the Regulation.[57]

3. Explainability in Article 13: an ambiguous and limited approach

From the realm of standardisation, "explainability" is the technical property of an AI system that refers to the relevant factors that influence

[56] Kiseleva, A., "Making AI's transparency transparent…", *Op. cit.*

[57] Cf. Recitals (1), (2), (3), (8), (25), (68), (102), (105), (119), (138), (139), (143), (146), Articles 1, 40.3, and Chapter VI of the Artificial Intelligence Regulation.

a decision and that can be expressed in a way that humans can understand. However, while explainability seeks to answer the question "Why?", it does not actually provide an argument that justifies whether the course of action taken was necessarily the most optimal.[58] For their part, the AI HLEG Ethical Guidelines define explainability as "the ability to explain both the technical processes of an AI system and the related human decisions (e.g., the application areas of a system)".[59]

In their case, explanations would be the means by which the decisions of an AI system can be explained in a clear, understandable, transparent and interpretable way for the addressee of the explanation. Therefore, if interpretability is the ultimate goal to be achieved, explanations are tools to achieve the interpretability of the model.[60] Consequently, "explainability" is an *active* characteristic of AI models that refers to their *technical* capacity to generate an explanation of their behaviour from the data used, the results obtained and the entire decision-making process[61], depending on the audience or profile of the addressees to whom the explanation is addressed.[62]

In particular, such an explanation must be timely and tailored to the level of expertise of the stakeholder (e.g., regulator, control authority, expert, researcher, affected by the decision, or the general public) in order for the system to be truly explainable.[63] In turn, explainable AI systems must comply with a number of basic principles: (i) that the system produces an explanation (by being inherently interpretable, technically on its own, or from comple-

[58] UNE-EN ISO/IEC 22989: 2023, 3.5.7. In a similar vein, IEEE Std 7001-2021 defines "explainability" as "the degree to which information made available to a stakeholder in a transparent manner can be readily interpreted and understood by a stakeholder".

[59] The Guidelines differentiate between *ad-intra* explainability ("technical explainability" vis-à-vis users or those responsible for the deployment of AI systems), and *ad-extra* explainability (collective or of the individuals concerned). "Technical explainability− explains the HLEG− requires that decisions made by an AI system can be understood and tracked by humans. Moreover, a choice may have to be made between increasing the explainability of a system (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a significant impact on people's lives, it should be possible to require an adequate explanation of the AI system's decision-making process. Such an explanation should be timely and tailored to the knowledge of the stakeholder (e.g. laypersons, regulators or researchers). In addition, it should be possible to obtain explanations of the extent to which an AI system influences and shapes the organisation's decision-making process, the system's design choices and the rationale for its deployment".

[60] Carvalho *et al.* (2019). "Machine Learning Interpretability…" *Op. cit.*, p. 15.

[61] iDANAE Chair (2019). *Interpretability*…, Op. cit., 3.

[62] Barredo *et al.* (2020), "Explainable Artificial Intelligence (XAI)… ". *Op. cit.*, p. 84.

[63] HLEG, *op. cit.*, paragraph 77.

mentary methodologies and metrics); (ii) that the explanation is meaningful and appropriate to the intended stakeholders; (iii) that the explanation reflects the system's processes accurately (as distinct from decision accuracy or predictive performance); (iv) and that the system expresses the limits of its design and domain.[64] Thus, while transparency would answer the question of how *the model works*, explainability would answer what *additional information can be extracted from the model* (explanations) when it is impossible or complex to see and understand (interpretability) how the model works internally (black box).[65] In turn, while interpretability would be the ultimate goal, explanations would be the tools to achieve interpretability when the model itself is not interpretable.[66]

For the European Commission's Expert Group, the British Data Protection Authority and the Alan Turing Institute, methods that include *post-hoc*, local or global XAI techniques (e.g., proxy models, Partial Dependency Graphs, LIME, Shapley Additive Explanations, Counterfactuals, etc.) are essential, not only to explain to users the behaviour of AI systems that are not inherently interpretable, but also to deploy a reliable technology.[67] Furthermore, explanations can include both technical measures (either *post-hoc* or explanations automatically generated by the system itself using AI technologies) and non-technical measures (e.g., written or spoken explanations in natural language about how the AI system works).[68]

A close look at the genesis, evolution and final text of the Regulation as a whole, and of Article 13 in particular, shows that explainability is not adequately addressed in the text. Moreover, despite the importance of this requirement in the Ethical Guidelines, explainability is only mentioned on three occasions: in the definition of "transparency" in Recital (48) in the terms explained above, in Article 13 and in Article 86 of the Regulation.

[64] Phillips, P. Jonathon, Hahn, Carina A., *et al.*, *Four Principles of Explainable Artificial Intelligence*, NISTIR 8312, 2021, https://doi.org/10.6028/NIST.IR.8312 The authors note that while established metrics exist to assess predictive performance, specific metrics to measure the accuracy of explanations are still in the process of being developed.

[65] Lipton, Z. C., "The Mythos of Model Interpretability". *ACM Queue*, vol. 16, no. 3 (2018), p. 12; Lepri, *et al.*, "Fair, transparent…", *Op. cit.*, p. 622.

[66] Kiseleva *et al.* "Transparency of AI in healthcare…" *Op. cit*, p. 6.

[67] HLEG, *Ethical Guidelines…*, *op. cit.*, paragraph 99; ICO and Alan Turing Institute, op. cit., pp. 120− 128, https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/ At the time of writing, the OECD has published a list of 16 specific explainability metrics. OECD.AI Policy Observatory, *Catalogue of Tools & Metrics for Trustworthy AI*, 2024. https://oecd.ai/en/catalogue/metrics?objectiveIds=11&page=1

[68] Kiseleva et al. "Transparency of AI in healthcare…" *Op. cit.* pp. 6-7.

If we go back to the legislative history, the simple fact is that, in the Commission's proposal, there was no trace of explainability throughout the entire text of the Regulation, with the sole exception of Recital (48).[69] On the other hand, interpretability did appear in paragraphs (1) and (3.d) of Article 13 in the same terms as in the final approved text. Now, in its final version, Article 13(3)(b)(iv) AIA provides that the instructions for use of high-risk systems shall incorporate:

> "where applicable, the technical capabilities and characteristics of the high-risk AI system to provide information that is relevant to *explain* its output [added italics]".

The current wording of Article 13(3)(b)(iv), which includes a very generic reference to "explainability", originates from Parliament's amendment No. 38.[70] In addition to Article 13(3)(b)(iv) of the Regulation, the other reference to explainability is contained in Article 86, which recognises the right to an *explanation* to the deploying officer of individuals affected by decisions taken by an Annex III high-risk system− but not of Annex I high-risk systems subject to harmonised legislation! The provision does not delimit what should be, in any case, the basic information to be provided to those affected in order to guarantee their right to a "clear and meaningful explanation [that can] provide as a basis on which the affected persons are able to exercise their rights".[71]

This being so, the AIA's approach to explainability has been the subject of different interpretations by the doctrine, which shows the ambiguity and lack of clarity of the legislator. Some consider that, as opposed to interpretability, explainability is in fact the Regulation's major shortcoming.[72] In this sense, it is

[69] Cf. Recital (48) of the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislation (COM/2021/206 final). The Recital included a minimal reference to the impact on the right to an effective remedy of high-risk systems used by public authorities for law enforcement purposes (investigation, detection, prevention and punishment of criminal offences by competent authorities) that are "not sufficiently transparent and explainable and not well documented".

[70] Amendment 308 of the European Parliament introduced a point iii.a (new) with the following text among the information to be included in the instructions for use: "the extent to which the AI system can provide an *explanation of* the decisions it takes [our italics]". Vid. European Parliament, P9_TA(2023)0236, *Op. cit.* Already, during the trialogue process, a version virtually identical to the current text of Article 13(3)(b)(iv) was included in the Interim Agreement of the Council, the Parliament and the Commission of 02/02/2024 (PE758.862v01-00).

[71] Recital (171).

[72] Kiseleva, A., "Making AI's transparency transparent…"; Schneeberger, *et al.*, "The Tower of Babel…". p. 70.

argued that Article 13 would not establish a general obligation of explainability for high-risk AI systems, but rather the transparency of the functioning of the system and the generation of results. In any case, this transparency should at least ensure that these elements are interpretable, which is not necessarily equivalent to the requirement of an explanation, at least in the terms explained above.[73]

Instead, other authors' reading of Article 13.1 in connection with Article 14.4(c) is that, through these specific provisions, the Regulation imposes on the provider an "explainability obligation" that is both enabling for the deployer and compliance-oriented. On the one hand, because such an obligation would serve to enable those responsible for the deployment of the AI system to interpret and use it correctly; on the other hand, because it would help to verify the compliance of the system with the obligations set out in the Regulation, ultimately contributing to achieving regulatory compliance.[74] However, we believe that the latter approach is misguided as it confuses interpretability with explainability.

It has also been understood that, by requiring in Article 13.1 "a sufficient level of transparency to enable deployers to interpret a system's output", the precept would cover both different types of explanations (local, global, counterfactual) and more or less granular information on the importance of variables. In any case, the explanations should be faithful to the model in the sense that they should be, at least approximately, a correct reconstruction of the internal decision parameters.[75] We do not agree with this reading of article 13 either, since it again confuses interpretability and explainability and, moreover, as will be argued below, the very wording of the precept seems to lean towards local explanations ("system output information").

In view of the above, the view held here is that the approach to the requirements of "interpretability" and "explainability" taken by the AIA in

[73] Bordt, S., Finck, M., Raidl, E., von Luxburg, U., "Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts". *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT 22)*, Association for Computing Machinery, New York, p. 894. https://doi.org/10.1145/3531146.3533153

[74] Sovrano *et al.*, "Metrics, Explainability…", *Op. cit.*, p. 132.

[75] Hacker, P. "Varieties of AI Explanations Under the Law…", p. 359. In a similar vein, see Onitiu, D., "The limits of explainability & human oversight in the EU Commission's proposal for the Regulation on AI- a critical approach focusing on medical diagnostic systems", *Information & Communications Technology Law*, vol. 32, no. 2(2022), p. 175. https://doi.org/10.1080/13600834.2022.2116354 However, Schneeberger, *et al*, "The Tower of Babel…", *op. cit.*, p. 70, are of the opinion that Article 13 leaves the interpretation open as to whether or not Article 13 AIA requires the application of post-hoc techniques and, if so, the approach to be chosen (local, global).

general and Article 13 in particular is ambiguous for the following reasons. Firstly, is not clear about the relationship/distinction between interpretability and explainability in Article 13. For a better argumentation and understanding of the provision, the following table incorporates the references to interpretability and explainability in Article 13.

Table 6. Interpretability and explainability in Article 13 AIA

Section	**Content**	**Purpose**
13(1.)	High-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to *interpret* a system's output and use it appropriately.	**Default interpretability [1]**
13(3.b.iv)	The instructions for use shall include, where applicable, the *technical capabilities* and *characteristics* of the high-risk AI system to provide information that is relevant to *explain* its output	**Explainability techniques [2].**
13(3.b.vii)	The instructions for use shall include, where applicable, *information* to enable deployers to *interpret* the output of the high-risk AI system and use it appropriately.	**Enabling information on interpretability [3].**
13(3.d)	The instructions for use shall include the human oversight measures referred to in Article 14, including the technical measures put in place to *facilitate the interpretation* of the outputs of the high-risk AI systems by the deployers.	**Interpretability techniques [4]**

Source: Own elaboration

When analysing the wording used by Article 13 in each of the sections indicated in the table above, one of the possible interpretations of how the provision integrates interpretability and explicability could be the following:

- In general, the provider should ensure that the design and configuration of high-risk systems are *interpretable by default* by the deployer [1].

- In case the model is not inherently interpretable (e.g., black box models), the instructions for use shall incorporate information on the techniques actually implemented by the provider to *generate explanations* that enable a proper interpretation of the output information (techniques and explainability tools) by the deployer [2].

- Although the model is interpretable by default, the instructions for use should contain sufficient information to *enable deployers to interpret* the output information and use the system correctly [3]. Such information may be useful when the users using the system implemented by the deployer are not AI experts (e.g., a doctor, a civil servant, or a worker within a private organisation), and even having some expertise the selected model requires mathematical or statistical tools to analyse decomposability and algorithmic transparency (see Table 4 *above*).

- Instructions for use should incorporate information on the human surveillance measures implemented[76], including the techniques put in place to *facilitate the interpretation* of output information from AI systems [4]. This could include information on the specific techniques that have been used to generate a default interpretable system, e.g., the type of interpretable model implemented by the provider (e.g., regression, decision trees, decision rule lists, K-nearest neighbours)[77]. But it could also refer to complementary techniques to generate explanations to enable the interpretation of the system results (e.g. LIME, surrogate interpretable models, partial dependency graphs, etc).[78]

[76] According to Article 14(4)(c) AIA, the human oversight measures implemented in the high-risk system shall enable the natural persons entrusted by the deployer with human oversight to "correctly interpret the high-risk AI system's output, taking into account, for example, the interpretation tools and methods available". The relationship of interpretability/explainability to human oversight measures, whether incorporated by the vendor at design (e.g., by incorporating appropriate human-machine interface tools) or implemented by the deployer on the vendor's recommendations, is also unclear. Given the wording of Article 14(4)(c) AIA it seems that the "tools and methods for interpretation" of system outputs would be part of the human oversight measures. And from the wording of the provision, "tools and methods of interpretation", the Regulation appears to refer to explainability techniques.

[77] See Information Commissioner's Office and Alan Turing Institute, "Explaining decisions…", *Op. cit.*, pp. 73-74. It is noted in the Guide that, "for AI models that are basically interpretable (such as regression-based systems, rule/decision lists, decision trees, Naïve Bayes or K nearest neighbour), the technical aspect of extracting a meaningful explanation is relatively straightforward. Typically, one resorts to the intrinsic logic of the model's mapping function through direct human observation […] For example, in decision trees or decision/rule lists, the logic underlying an outcome will depend on the interpretable relationships of the weighted conditional (if-then) statements. In other words, each node or component of such models functions, in effect, as a reason […] In general, it is useful to be aware of the *range of techniques available* for building interpretable AI models […]. These techniques not only make the foundations of AI models easily understandable, but also form the basis of many of the complementary explanatory tools that are widely used to make 'black box' models more interpretable." That is, interpretable models ("technical measures established to facilitate the interpretation of output information from high-risk AI systems") can be used, for example, as surrogate models to explain the results obtained by non-interpretable models.

[78] See *ibid.* The Information Commissioner's Office and Alan Turing Institute Guidance notes that if, after considering domain, impact and technical factors, a "black box" AI system

In addition to the above, from the wording used by Article 13, it seems to suggest that only a *local level of interpretability* would be required, since in its various paragraphs the interpretability requirement is limited exclusively to the "system output information", thus excluding other elements of the system, such as the model and its components (variables, parameters, interactions, processing algorithm).[79] This local interpretability requirement could be interpreted in the sense that the Regulation would be prioritising local explainability techniques, rather than global explainability techniques.

However, while local explanations are critical in cases where system decisions impact on individual people, global explanations make it possible to understand the relationship between system components and their behaviour as a whole. In this sense, global explanations will often be critical not only to establish an accurate local explanation, but to ensure the fairness, safety and optimal performance of your AI system. In addition, a global understanding of the system can also provide essential information about the broader potential impacts of the system on specific groups and society at large.[80] In this sense, global explanations could be relevant to explain the operation of the high-risk system "regarding specific persons or groups of persons on which the system is intended to be used" (see Article 13.3.b.v), and even identify known or foreseeable circumstances, associated with "the use of the high-risk AI system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to risks to the health and safety or fundamental rights" of individuals and to society as a whole (see Article 13(3)(b)(iii)). However, given the vagueness of the Regulation, providers could rely on a restrictive interpretation based on local interpretability as a *de minimis* criterion of relying on a literal interpretation of Article 13.

has been chosen, appropriate supplementary explanation tools should be incorporated into the construction of its model. Complementary explanation strategies available according to the state of the art to support interpretability "can shed light on significant aspects of a model's overall processes and components of its local results".

[79] For the components of an AI model, see ISO/IEC DIS 12792: 2024 (en), 9.

[80] Local interpretations aim to interpret individual predictions or classifications corresponding to specific instances in order to identify the specific input variables that may have been determinant or have had more weight in the generation of a particular prediction or classification. Global explanations, on the other hand, seek to offer a broad view that encompasses the general importance of the variables and their interactions in the results generated by the model, the inner workings and the logic of the behaviour of the model as a whole. Global interpretations focus on explaining the model as a whole, rather than behaviour for a particular case, and can contribute to a procedurally coherent decision-making process. See Information Commissioner Office's, Alan Turing, *Explaining decisions… Op. cit.*, p. 74.

IV. Subjective and formal scope of Article 13

Paragraphs (1) and (2) of Article 13 delimit the transparency requirement on the basis of two areas. First, it identifies the obliged entities and recipients of the transparency requirement for high-risk systems. Second, it sets out the basic formal aspects that modulate compliance with the obligation, namely the level of transparency required, as well as the manner in which the relevant information must be communicated to the deployer.

1. Subjects and the purposes of transparency in Article 13: the regulation's major absences

There is common agreement in the doctrine that the transparency requirement set out in Article 13 is limited to two specific subjects: the provider, who is bound by the transparency requirement, and the deployer, who is the recipient of the information provided for in Article 13.3.

As anticipated above, Article 13.1 sets out a type of internal transparency. The provider, according to Article 3.3 AIA, can be any "natural or legal person, public authority, agency or other body that *develops* an AI system or a general-purpose AI model or that *has* an AI system or a general-purpose AI model *developed* and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge [added italics]".

This concept would thus cover circumstances where the provider is directly responsible for the design and development of the system, as well as those where a third party has designed and developed the system for the provider, with the latter being responsible for its introduction on the market or putting into service under its own brand name or trademark. In turn, the recipient of the information contained in Article 13.3 AIA is the person responsible for the deployment, which according to Article 3.4 would be identified as any "natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity".

In the case of deployers, the transparency obligation in Article 13 would have a dual (compliance and enabling) purpose. On the one hand, transparency enables the regulatory compliance of deployers with the obligations under Section 3 of Chapter III, and in particular with the obligations under Article 26 (inter alia, human oversight, monitoring of the operation of the high-risk AI system, preservation of log files automatically generated by high-risk AI systems, or compliance with the impact assessment). On the other hand, transparency would also enable deployers to be able to interpret the output information and to use the system correctly in accordance with the instructions for use provided by the provider (Article 13.1).

In the case of providers, the purpose of transparency is regulatory compliance (Article 13.1) and, where applicable, accreditation of such compliance to the supervisory authorities. In the first case, regulatory compliance refers to the obligations set out in Article 18, including compliance with the requirements defined in section 2, implementation of the quality management system, documentation keeping, adoption of corrective measures and their communication, retention of the logs automatically generated by their systems, compliance with the conformity assessment procedure, drawing up of the declaration of conformity, registration of the high-risk system in the EU database.

Alongside regulatory compliance, transparency (through communication and traceability) also enables the provider to demonstrate regulatory compliance to the competent national authority of the compliance of the high-risk AI system with the requirements set out in Section 2 of Chapter III (Article 18.2(K)).

In contrast to this internal transparency, one of the most unanimous criticisms of Article 13 and, in general, of the entire Regulation, is that the approach to transparency is exclusively technical and strongly limited from a subjective point of view. And, therefore, the AIA would in no case be truly enabling for the exercise of rights by those affected by high-risk schemes, as no clear framework is established to provide individuals with clear avenues to challenge decisions taken by AI schemes that affect them.[81] Where appropriate, Article 13 would have established a sort of "expert transparency for experts", exemplified by the list of information described by Article 13(3) and to be included by the provider in the instructions for use, which are targeted and addressed exclusively to the deployer. In this sense, the Regulation would have established a particular and restricted objective of transparency, which would be limited exclusively to facilitating compliance by providers and deployers with the obligations set out in Section 3 of Chapter 3 (Article 16-27), to the detriment of transparency aimed at the exercise of rights by the persons affected by the decisions of high-risk systems. In particular, Article 13 would have set up "a novel type of instrumental, self-referential and compliance-oriented transparency focused on the effective and compliant implementation of AI systems in specific environments".[82]

[81] Onitiu, D., "The limits of explainability…", *Op. cit.*, p. 171; Smuha, *et al.*, "*How the EU Can Achieve Legally Trustworthy AI*", *Op. cit.*, p. 52.

[82] Hacker, P., Passoth, JH., "Varieties of AI Explanations Under the Law…", *Op. cit.*, p. 361.

This restrictive approach to transparency would be in clear contradiction with the approach of the explanatory part of the AIA, where it is continuously stressed, in one way or another, that the essential objective of the Regulation is "to promote the uptake of human centric and trustworthy Artificial Intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (the 'Charter'), including democracy, the rule of law and environmental protection".[83]

2. The formal scope of transparency: the undefined "appropriate type and level of transparency".

According to Article 13.1 AIA, "[h]igh-risk AI systems shall be designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately. *An appropriate type and degree of transparency* shall be ensured with a view to achieving compliance with the relevant obligations of the provider and deployer set out in Section 3 [added italics]"

The approach adopted by the AIA is ambiguous: neither the type nor the level of transparency that is considered adequate is specified.[84] The appropriate form and level of transparency appear to be relative and merely instrumental to achieving compliance with other requirements of the AIA.[85]

It is possible that the technical standards to be developed by CEN and CENELEC in response to the request for standardisation made by the European Commission will develop this issue[86] or, failing that, the common specifications that may be adopted by the Commission by means of implementing acts.[87]

In the meantime, when determining the type and level of transparency required, it should be noted that the different technical standards consulted (ISO/IEC DIS 12792:2024(en) or IEEE 7001-2021) establish the need to

[83] Vid. Recital (1) of the Regulation. And, in a similar sense, see Recitals (8)-(10), (20), (28), (32), (43), (46), (48) among many others.

[84] See. Onitiu *et al.*, "The limits of explainability…", *Op. cit.*, p. 174; Kiseleva, "Making AI's Transparency transparent…", *Op. cit.*; Boch, A., Hohma, E., Trauth, R., *Towards an Accountability Framework for AI: Ethical and Legal Considerations*, Technical University of Munich, Munich Center for Technology in Society, Institute for Ethics in Artificial Intelligence, February 2022, p. 5.

[85] See Schneeberger*, et al.*, "The Tower of Babel… ", *Op. cit.*, pp. 65, 70; Hacker, P., Passoth, JH., "Varieties of AI Explanations…", Op. cit., p. 359.

[86] Vid. Article 40 AIA and Commission Implementing Decision of 22.5.2023 on a standardisation request (C(2023) 3215 final), *Op. cit.*

[87] Vid. Recital (121) and Article 41 AIA.

tailor the type, level, and even format of relevant information to different stakeholder profiles.

For the purpose of this Chapter and for determining the meaning and scope of the term "type and levels of transparency", the standard IEEE 7001-2021 for autonomous systems[88] is interesting because it precisely graduates the required levels of transparency from 0 (lowest) to five (highest)[89] according to the role of the stakeholders involved (system users, general public, certification or regulatory bodies, incident/accident investigators, and expert advisors in administrative or judicial proceedings).

In turn, the rule distinguishes between different sub-categories of system users that would require different levels of transparency:

- Non-expert users, which include both people who have only brief interaction with the system and people who interact frequently with the system.

- Domain expert users include users with knowledge and experience in the domain in which the system is applied (e.g., a doctor). These users have some responsibility for the use of the AI system.

- Super-users are experts not only in AI systems, but also in the specific systems for which they are responsible. These super-users include people responsible for the development, fault diagnosis, repair, maintenance and upgrading, as well as the operation and monitoring of specific autonomous systems.

The following table incorporates the different levels of transparency according to user profile. Note that, from transparency level 3 onwards, the standard establishes different explainability requirements appropriate to the user profile.

[88] The scope of the standard covers all autonomous systems, both physical and non-physical. The former include vehicles with automated driving systems or assistive robots, while the latter include medical diagnostic systems (recommenders) or chatbots. Intelligent autonomous systems using machine learning also fall within the scope of the standard. Likewise, the datasets used to train such systems are also within the scope of the standard when considering the transparency of the system as a whole. See IEEE 7001-2021, 1.1.

[89] Within each stakeholder category, requirements are set for measurable and verifiable levels of transparency. Transparency levels are defined from 0 (no transparency) to 5 (the highest achievable level of transparency). Each definition is a requirement expressed as a qualitative property of the system that must be met. Levels 1 to 5 have been defined to describe successively higher levels of transparency. All levels are considered technically feasible, while each successive level is generally more demanding. Each level is cumulative and builds on the previous ones, so that when a system meets level n of a particular category, it is expected that it will also meet levels n -1. In each case, verification of the level is simply a matter of determining whether or not the requirement is met, i.e., whether or not the transparency property required by a given level for a given stakeholder group is demonstrably present. *Idem,* 5.

Table 7. Levels of transparency and explainability for users in IEEE 7001-2021

Level	**Definition**
0	There is no transparency.
1	Accessible information shall be provided to the **user**, including at least: (a) example scenarios with expected and intended system behaviour, including degraded modes of operation; and (b) general principles of operation, i.e., whether there is a learning component and what data it uses. The documentation should explain the general principles of operation of the system. In the case of a system using machine learning, the documentation should provide a simple explanation of what sources the system examines/uses as part of the learning process, including potential sources of bias. This documentation could, for example, take the form of a written manual, pictorial guide or audio guide, depending on the user's needs, explaining how the system behaves in the different circumstances and situations that its designers expect it to encounter. Expert **users and super-users** shall receive the user documentation specified above and prepared in accordance with IEC/IEEE 82079-1. This documentation shall detail the safe operation and monitoring of the system. For **super-users**, the documentation shall detail procedures for system fault diagnosis, repair, maintenance, upgrade and decommissioning at end-of-life.
2	The **user** will be provided with interactive training material to allow them to test their interactions with the system in specific and relevant virtual situations. In addition, **expert users and super-users** will receive interactive training material on the safe operation and monitoring of the system. Super-users will also receive interactive training material on fault diagnosis, repair, maintenance, upgrade and end-of-life decommissioning.
3	The **non-expert user** shall be provided with user-initiated functionality that produces a brief and immediate *explanation* of the most recent system activity. These *explanations* shall be expressed through commonly understandable means, such as natural language or other appropriate means (e.g., an image). Neither the making of requests nor the understanding of the system's responses to such requests shall require any training of the non-expert user. However, warnings for security or legal reasons that are necessary shall be accepted. For systems designed for use by **expert users**, the same functionality specified above shall be provided, except that (a) the system shall allow *explanations* of any of its recent decisions to be requested and (b) the *explanations* shall be expressed using language appropriate to the subject matter. In addition, experts shall be provided with documentation detailing how these explanations are to be requested and interpreted. Such documentation shall also include natural language processing (NLP) subsystems, if any.

4	The non-expert user will be provided with user-initiated functionality that produces a brief and immediate explanation of what the system does in a given situation. Compliance with this level of transparency allows the user to explore hypothetical "what-if" scenarios in a given situation, if applicable to the system's scope of work. Neither making requests nor understanding the system's responses to such requests will require the non-expert user to receive any training, although familiarity with the system's user documentation is necessary. For systems designed for use by **expert users**, the same functionality specified here shall be provided, except that *explanations* may be expressed using language appropriate to the subject matter. In addition, documentation shall be provided to expert users detailing how these *explanations* are to be requested and interpreted. Such documentation shall also include NLP sub-systems, if any. Importantly, this level of transparency allows the user to extract *counterfactual explanations*.
5	The **user** shall be provided with a continuous behavioural explanation that adapts the content and presentation of the *explanation* according to the user's information needs and context. This shall include access to log files and training data as long as they do not contain sensitive information such as personal data. *Explanation* of operation shall be achieved by some simple and visible visual presentation, after the system performs an action, or by vocalisation of explanatory phrases while the system performs an action. Non-expert users shall not be required to make additional effort to access the relevant explanations. This interaction should be tailored to the user's interaction history, as trust is easily lost if, for example, the system behaves unexpectedly. Additional explanatory details will be available on request, as required by **expert users** or **super-users**, allowing them to interactively explore the system and its operation.

Source: IEEE 7001-2021, 5.1.1

In addition to requiring types and levels of transparency appropriate to regulatory compliance by the provider and the deployer, Article 13.2 of the Regulation prescribes certain formal requirements for the relevant information to be incorporated in the instructions for use accompanying the system: "concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers." In this regard, transparency should take into account the possible perception and understanding of stakeholders and, where appropriate, avoid disclosing information in a way that, while technically true, is framed in a way that leads to misinterpretation.[90]

[90] IEEE 7001-2021, 3.1.

V. The material content of the transparency obligation

From a material point of view, the provider of the high risk system should include in the instructions for use provided to the deployer the information referred to in Article 13.3 AIA. Firstly, it is important to clarify that the instructions for use cannot be confused with the technical documentation (Article 11) which the provider is required to keep for a period of ten years at the disposal of the competent national authorities (Article 18.1(a)).[91] The clarification does not go without saying. Secondly, Article 13.3 does not contain a *numerus clausus* of categories of information, but should be interpreted as a *de minimis* criterion, as the Regulation is clear in stating that "[t]he instructions for use shall contain *at least* the following information […] [added italics]".

The instructions for use provided to the person responsible for deployment shall include at least the following information:

- The identity and the contact details of the provider and, where applicable, of its authorised representative (Article 13(3)(a)).

- The characteristics, capabilities and limitations of performance of the high-risk AI system, including (article 13(3)(b)).

- Changes to the system and functional suitability predetermined by the provider (Article 13(3)(c)).

- The human oversight measures envisaged, including techniques to facilitate the interpretation of output information from the system (Article 13(3)(d)).

- The computing resources, hardware and expected lifetime of the system (Article 13(3)(e)).

- Logs implemented (Article 13(3)(d)).

Apart from the fact that nothing in Article 13 or elsewhere in the Regulation suggests that the provider cannot extend this information if it considers that this would contribute to increasing the transparency of the system with a view to improving the capability of the deployer, it will have to await the technical standards to be developed by CEN and CENELEC as to whether this *de minimis* information can be extended with other relevant information. In this respect, it should be noted that ISO/IEC DIS 12792:2024(en) includes with-

[91] Vid. Article 11(1) AIA: The technical documentation for a high risk AI system shall be drawn up before it is placed on the market or put into service and shall be kept up to date. The technical documentation shall be drawn up in such a way as to demonstrate that the high risk AI system complies with the requirements laid down in this Section and shall provide the national competent authorities and notified bodies with the information necessary to assess the conformity of the AI system with those requirements in a clear and comprehensive manner. It shall contain at least the elements set out in Annex IV.

in each level of transparency taxonomies categories of information which are not covered by Article 13, and which could be adapted to the profile of the stakeholder to whom the information is addressed.[92]

Thirdly, Article 13.3 does not identify the level of detail with which the provider must specify the various categories of information identified in sub-paragraphs (a)-(f). In fact, those categories of information included in Article 13.3 coincide, in turn, with many of the categories of information set out in Annex IV which form part of the content of the technical documentation.[93] It is not clear, however, whether the level of detail of the information contained in the instructions for use should be qualitatively and quantitatively lower than that of the technical documentation, taking into account, moreover, the different recipients and purposes of communication of one or the other information.[94]

In addition to the above, paragraph (3) of Article 13 uses limiting phrases ("when appropriate", "where applicable") which, on a literal reading, could lead to restrictive interpretations of some provisions. This would be the case of paragraphs (3)(b)(vi) (in relation to data used by the system), (3)(b)(vii)

[92] Thus, for example, the context level could include relevant information regarding the environmental impact of the system, e.g., environmental impact assessments performed, the energy consumption of the system, its carbon and water footprint, system decommissioning, and waste management. The system level includes access to the internal elements of the AI system, including certain individual components; parts of the source code; model-specific elements (list of rules or knowledge elements embedded in graphs, the parameters in machine learning models), and internal and intermediate values resulting from the processing of a particular input. At the model level, access could include the model's dependency on other models, the concrete processing algorithm, the procedure to build the model, the hyperparameters, or the input and output data formats. Last but not least, at the data level, important details could include where the data came from, its statistical properties, its biases and limitations, how it was collected and prepared, how it was labelled, how imbalances were found, and what anonymisation and pseudo-anonymisation steps were used.

[93] See Table 2 of correspondence between Article 13 AIA, Articulated AIA and Standardisation, including equivalences between the sub-paragraphs of Article 13(3) and the categories of information in Annex IV.

[94] In the case of instructions for use, the addressees are the deployers and their purpose is the regulatory compliance of the deployers with the obligations laid down in Section 3 of Chapter III, as well as their training to enable them to interpret the output information and to use the system correctly (Article 13(1)). In the case of technical documentation, the ultimate recipients are the national competent authorities and notified bodies and its purpose is to enable authorities and bodies to assess the conformity of the AI system with the requirements set out in Section 2 of Chapter III of the Regulation (Article 11(1)), including, therefore, the transparency requirement.

(proper interpretation of output information and correct use of the system), or (3)(f) (in relation to logs).[95]

Finally, if we take into account the taxonomies of transparency levels identified in ISO/IEC 25059:2023(en) (context level, system level, model level, data level), it can be seen that most of the information categories foreseen in Article 13.3 refer to the system level (intended purposes, capabilities, functional limitations, recommended and prohibited uses, log preservation and archiving, human factors), and only some of them to the model level (predictive performance, computational and hardware resources), and to the data level (basic information).[96]

1. Information on functional suitability and other properties

Article 13.3(b) of the Regulation provides that the instructions for use accompanying high-risk systems must include information on "the characteristics, capabilities and limitations of performance of the high-risk AI system". The term "performance of an AI system" is to be understood as meaning in the sense indicated by the Regulation itself "the ability of an AI system to achieve its intended purpose" (Article 3.18). This concept coincides with the technical notion of "functional suitability", i.e., "the ability of the system to provide functions that facilitate the accomplishment of specified tasks and objectives".[97] In turn, the "intended purpose" of the high-risk system (cf. Article 13.3(b)(i)) would be "the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documen-

[95] The grammatical restriction introduced in the above paragraphs means that, "when appropriate" or "where applicable", the provider might not provide the information concerned, which would conflict with the very purposes of Article 13 with regard to deployers. In some cases, because information would be omitted which could be relevant for the proper interpretation of the output results and the correct use of the system. In others, because the information in question enables the controller's regulatory compliance with some of the obligations imposed by Article 26 AIA, inter alia, human supervision (paragraph 2); the relevance and representativeness of the data used by the system, especially if the control of the data is with the provider (paragraph 3); the monitoring of the functioning of the high-risk system (paragraph 5); the retention of log files (paragraph 6); and even the proper implementation of the data protection impact assessment where this is required (paragraph 9).

[96] On the four levels of transparency taxonomies in ISO/IEC 25059:2023(en), see Table 2 *above*.

[97] Regarding functional suitability, see ISO/IEC 25010:2023 (en), 3.1; ISO/IEC 25059:2023 (en), 5.1, Figure 1.

tation" (Article 3.12). Here, the concept of "performance" is identified with the purpose of the system (its ability to produce a prediction, a recommendation, a decision), intended, stated, documented and tested by the provider for a specific context or domain and specific implementation conditions.

In identifying the intended purpose of the high-risk system (Article 13.3(b)(i)), the instructions for use should describe what objectives of the system according to the user's needs it can address and how AI can contribute to achieving those objectives.[98] Given that the purpose depends, among other factors, on the specific context of the high-risk AI system, it seems relevant that the instructions for use incorporate basic information on the social context (intended geographical location of implementation, socio-occupational or organisational context of deployment, linguistic or cultural constraints of the system).[99]

Information on the social context may, in turn, be relevant to the deployer's ability to correctly interpret any "risks to health and safety or fundamental rights" anticipated or resulting from misuse (Article 13.3(b)(iii)), or the operation of the system "regarding specific persons or groups of persons on which the system is intended to be used" (Article 13(3)(b)(v)).

In any case, it should be noted that, during the legislative process of the Regulation, the text presented by the Council as opposed to the Commission's proposal envisaged the addition of a subparagraph to paragraph (3)(b)(i), which would include not only the intended purpose, but also information on "the specific geographical, functional or behavioural environment in which the high-risk AI system is intended to be used". However, this approach was not incorporated in the final text of the Regulation.

In addition to functional adequacy, Article 13.3 provides for several categories of information that relate to other technical properties of AI systems, such as robustness and safety, or performance efficiency. With regard to robustness and safety, these are two requirements that, together with accuracy, high-risk systems must meet under Article 15. With regard to robustness and safety, Article 13.3(ii) refers to the "tested and validated level of the high-risk AI system that can be expected, as well as any known and foreseeable circumstances that may have an impact on that expected level". However, the provision does not define what specific quantitative or qualitative information

[98] Cfr. ISO/IEC DIS 12792:2024 (en), 8.4.2.

[99] Cfr. ISO/IEC DIS 12792:2024 (en), 7.2.1. ISO/IEC 22989:2022 (en), not only considers social impact (5.18), but also aspects related to jurisdictionality (5.17), as in the country where the system has been designed or produced it might be subject to different legal requirements than in the European Union.

should be provided to the deployer to meet the transparency requirement (interpretation of output results, correct use of the system and compliance with Article 26 obligations). For example, Article 13 explicitly includes information on performance metrics, as discussed below, but not on robustness metrics, which, on the other hand, are explicitly mentioned in Article 15. Similarly, the basic information to be included in the instructions for use on technical and organisational security measures to meet the transparency purposes of Article 13 is also unclear.

Some paragraphs of Article 13 also include information relating to "performance efficiency". This property represents the ability of the system to perform its functions within specified time and performance parameters and to be efficient in its use of resources under specified conditions. Resources can be CPU, memory, storage, network devices, other software products with which the system interacts, or energy used.[100] In this regard, Article 13.3(e) AIA also foresees that information regarding the "computational[101] and hardware resources required, the expected lifetime of the high-risk AI system, as well as the maintenance and care measures required, including their frequency, to ensure the proper functioning of that system, including with regard to software updates, shall be provided to the deployer".[102]

2. Information on functional correctness or predictive performance: "accuracy" and its "metrics".

In AI systems, the so-called "functional correctness" or "predictive performance"[103] is one of the technical properties that can have the greatest

[100] ISO/IEC 25010:2023(en), 3.2. See also, Janapa Reddi, Vijai (ed.) *Machine Learning Systems with TinyML*, Harvard University, last updated 21 March 2024, p. 392. https://harvard-edge.github.io/cs249r_book/contents/benchmarking/benchmarking.html

[101] In the Spanish version of the Regulation, the term "recursos informáticos" is used instead of "recursos *computacionales*" ("*computational* and hardware resources").

[102] Note that ISO/IEC DIS 12792:2024 (en), 9.4.7 and 9.4.8 includes in the transparency taxonomy corresponding to the model level the type of computing hardware and computational costs (e.g. total CPU and GPU time per input data sample or per input data size).

[103] ISO/IEC 25059:2023(E), Annex C. While it is common in the field of AI to use the term "predictive performance" to mean how well a particular AI system performs its intended tasks, the standard clarifies that it is preferable to refer to this property as "functional correctness" to clearly differentiate it from "performance efficiency". Predictive performance" refers to the generalisability of the model, i.e., its ability to obtain accurate results with new and unknown data inputs, beyond the specific examples with which the model was trained. Vid. Martínez-Heras, Jose Antonio, IArtificial.net. Technical Report, 2023. DOI: 10.13140/RG.2.2.16587.77609M; Ministry for Digital Transformation and Red.es, How do I know if

impact on the reliability of AI systems[104] and thus on the interpretation of the output results and their correct use by the deployer.

Functional correctness' defines the ability of the system to provide correct results with the required degree of accuracy. AI systems, and in particular those using machine learning models, do not, however, usually provide functional correctness in all observed circumstances because a certain error rate is expected in their results. Therefore, there are numerous metrics that assess functional accuracy. In addition, depending on the context of use and the purpose of the system, trade-offs between functional correctness and other system properties, such as performance efficiency or robustness, may be necessary.[105] Moreover, functional correctness could also be affected by cybersecurity. For example, the European Cybersecurity Agency identifies among the threats to learning models, the compromise of inference correctness; the reduction of the level of accuracy of data by modifying or mixing it with other datasets of different qualities; or the manipulation of labelled data in supervised models[106]. In turn, the deployment of security controls often leads to a delicate balance between system security and system performance.[107]

Functional correctness– in addition to other properties, such as robustness and cybersecurity– seems to be referred to in Article 13.3(b)(ii), by requiring that the instructions for use incorporate specific information on the "*level of accuracy (including the parameters for assessing it)*, robustness and cybersecurity referred to in Article 15 against which the high-risk AI system has been tested and validated and can be expected, as well as the known and foreseeable circumstances that could affect the expected level of accuracy, robustness and cybersecurity [added italics]." In turn, Article 15.3 of the Regulation reiterates this provision, stating that "[t]he instructions for use accompanying

my prediction model is really good?, 26 January 2020. https://datos.gob.es/en/blog/how-do-i-know-if-my-prediction-model-really-good In relation to machine learning models, ISO/IEC 42001:2023 defines "generalisation" as "the ability of a trained model to make correct predictions from previously unseen input data".

[104] Cf. ISO/IEC 22989:2022 (en), 5.15.3. Precisely, the standard defines reliability in AI systems as the capability of the system that "enables it to provide the required prediction […], recommendation and decision *consistently and correctly* during its operational phase [emphasis added]".

[105] ISO/IEC 25059:2023(E), 3.2.3, 5.4.

[106] ENISA, *AI Cybersecurity Challenges. Threat Landscape for Artificial Intelligence*, December 2020, DOI 10.2824/238222, p. 44-47.

[107] See ENISA, *Securing machine learning algorithms*, December 2021, pp. 3, 27. DOI: 10.2824/874249.

high-risk AI systems shall indicate the levels of accuracy of such systems, as well as the *relevant parameters for measuring accuracy*".[108]

With regard to the "level of accuracy (including its metrics)" both the expression used in the Spanish version and in the English version ("nivel de precisión (incluidos los parámetros para evaluarla") two preliminary remarks should be made. First, the Spanish translation of "metrics" as "parámetros para evaluar [la precisión]" is inappropriate.[109]

Firstly, because the term "parameter" has a specific technical meaning. And it is in such a technical sense that this term is used in Recitals (98, 102, 104), referring to general purpose models and, in particular, to "weights", or in Article 3, paragraphs (29) and (30), in relation to trainable and non-trainable parameters (or hyperparameters). Secondly, because the term 'accuracy' in a narrow sense refers to a specific performance metric of LM-based classification models.[110] From a systematic interpretation of Article 13.3(b)(ii) AIA in connection with the expository part of the AIA, it should be concluded that the term "level of accuracy (including the parameters for assessing it)" necessarily refers to the identification and description in the instructions for

[108] Again, the Spanish version of the Regulation translates the term "métricas" as "parámetros pertinentes para medir[…] [la precisión]".

[109] Functional correctness or predictive performance is a measurable and assessable property, quantitatively and qualitatively (ISO/IEC 42001:2023, 3.11; ISO/IEC 25059:2023(en), Annex C) by means of so-called "performance metrics", also called "error metrics" or "evaluation metrics". These metrics include logical-mathematical constructs designed to measure the closeness or closeness between the predicted outcome (prediction) and the actual outcome. In other words, performance or error metrics allow an assessment of the quality of the model in terms of its predictive capacity. In this sense, the greater the difference between the actual outcome "r" and the predicted outcome "p", the more "distant" the model is from being an accurate representation of reality. Conversely, the closer the estimated values "p" are to reality "r", the better the model will perform in predictive terms. Vid. Plevris, Vagelis; Solorzano, G.; Bakas, N. P.; et al, "Investigation of performance metrics in regression analysis and machine learning-based prediction models", *ECCOMAS Congress 2022 - 8th European Congress on Computational Methods in Applied Sciences and Engineering*, 2022, https://www.scipedia.com/public/Plevris_et_al_2022a.

[110] The accuracy metric measures the percentage of cases (true positives and true positives) that the model has been correct. ISO/IEC 23053:2022(E), 6.5.5.4, identifies the most common metrics for classification (*accuracy*, precision, confusion matrix, recall, F1 score) and regression models (mean absolute error, root mean squared error, relative absolute error, relative squared error, mean zero one error, coefficient of determination). On the application of error metrics to AI solutions procured by the National Health System domain, see Gutiérrez David, M.E. and Quintana Cortés, J.L., "Public Procurement of AI for the EU Healthcare Systems. First Insights from the Spanish Experience", *European Review of Digital Administration & Law - Erdal*, Volume 4, Issue 1 (2023), pp. 131-132.

use of the functional correctness or predictive performance of the model, as well as of the performance or error metrics implemented in the AI system by the provider.

On the provider side, measuring the predictive performance of an AI model can serve different purposes:[111]

- The evaluation of the model, to find out how reliable its predictions are[112] or the frequency and expected size of its errors [1].

- The comparison of different models, in order to choose those with the best trade-offs between performance and efficiency[113] [2].

- Out-of-sample comparisons and over time, to check that model performance has not degraded with new production data[114] [3].

- The determination, depending on the use case and application context, of the most optimal way to compensate for the (usually inverse) relationship between the performance of the model and its level of interpretability [4].

- The design of more interpretable and, where appropriate, explainable models, while maintaining high levels of performance [5].

Providing the deployer with relevant information on the predictive performance of the system contributes to ensuring the enabling purpose of transparency in Article 13 (correct interpretation of output results and proper use of the system) and regulatory compliance in the following terms.

Firstly, providing relevant information on the functional correctness of the system, the frequency, or size of its errors, contributes to improving the

[111] See, ISO/IEC TS 4213:2022*,* 6.

[112] Liu, Zhenyu and Chen, Huanhua, "A predictive performance comparison of machine learning models for judicial cases", in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, Honolulu, 2017, pp. 1-6, DOI: 10.1109/SSCI.2017.8285436. For example, the authors have evaluated and compared the performance of different machine learning algorithms (k-NN, logistic regression, bagging, random forests and support vector machines) in predicting judicial decisions based on a selection of variables representing the semantic context of cases from the HUDOC database of the European Court of Human Rights, alleging violations of Articles 3 (prohibition of torture and inhuman or degrading treatment), 6 (right to a fair trial) and 8 (right to respect for private and family life, home and correspondence).

[113] For example, in Deng, Fei, Huang, Jibing, Yuan, *et al.* "Performance and efficiency of machine learning algorithms for analysing rectangular biomedical data", in *Laboratory Investigation*, vol. 101, 2021, pp. 430-441, https://doi.org/10.1038/s41374-020-00525-x, the performance of different machine learning models (decision trees, random forests, support vector machines and artificial neural networks) for multi-category classification of causes of death (survival, breast cancer, other cancers, cardiovascular diseases, other causes) from large biomedical datasets is comparatively analysed.

[114] Janapa Reddi, Vijai (ed.) Machine Learning Systems with TinyML, Harvard University, last updated 21 March 2024, p. 607. https://harvard-edge.github.io/cs249r_book/contents/benchmarking/benchmarking.html

reliability of predictions. The reliability of predictions depends not only on the implementation of appropriate error metrics according to the context and intended purpose of the system, but also on the quality of the data used to train, validate and test the models. It is therefore relevant for Article 13.3(vi) in the instructions for use to incorporate "specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the high-risk AI system". Adequate information about the data and its pre-processing can help to detect inherent biases in the data and sources of potential discrimination.

Secondly, given the evolutionary nature of some AI systems, it is imperative to test the behaviour of the system over time with new production data. It is therefore appropriate for Article 13.3(c) to include in the instructions for use "changes to the high-risk AI system and its performance which have been pre-determined by the provider at the moment of the initial conformity assessment". This could include a description of the mechanisms implemented to ensure that the behaviour of the model evolves as intended within the version predetermined by the provider.[115]

In addition, new risks to the persons affected by the system not initially foreseen by the provider may arise during the lifecycle of the AI system due to certain uses by the deployer, whether fit for purpose or inappropriate. Such risks could be due to the input of new data with a different distribution and representativeness than those used to train, validate and test the model. Hence, Article 13.3(b)(iii) provides for the incorporation in the instructions for use of information on "any known or foreseeable circumstance, related to the use of the high-risk AI system in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, which may lead to risks to the health and safety or fundamental rights".

Third, although already explained *above,* knowing the degree of interpretability and explainability of the system while maintaining it is relevant in critical contexts where an error rate above a certain threshold or the absence of an adequate level of interpretability and explainability has adverse consequences for health, safety or fundamental rights. [116]Hence, Article 13.3(iv)

[115] ISO/IEC DIS 12792:2024(en), 9.4.9. According to the standard, such mechanisms could include the existence of databases that aggregate new information for the use of the (unmodified) model; the use of production data to modify the model in real time; the storage and exploitation of deployment side operations (e.g. model decision correction) or other forms of feedback to influence or modify the behaviour of the model.

[116] For example, in critical contexts such as healthcare or criminal justice. Cf. Rudin, Cynthia, "Stop Explaining Black Box…", *op cit*, pp. 206-207.

includes information relating to "specific persons or groups of persons on which the system is intended to be used".

VI. Final assessment of Article 13 of the Regulation

This is followed by a final assessment of article 13 of the AIA regarding its improved drafting technique and the possible discouragement of black box models, as well as the lack of definition of the "appropriate type and level of transparency". It also assesses the "de minimis" content of material transparency, without specifying its scope, the standardisation and transparency of high-risk systems and, finally, the "great forgotten ones" of the AIA.

1. Improved drafting technique

The systematic followed and the content of Article 13 contain some limitations which may give rise to interpretative complexity:

- Explicit and implicit references to other provisions of the Regulation or to specific technical terminology to be integrated in the interpretation of Article 13.

- Inappropriate use of technical concepts or poor translation of technical concepts (at least in the Spanish version (performance/functioning, metrics, parameter, precision).

- Use of ambiguous and open expressions ("type and levels of transparency") or the lack of determination of the degree of detail with which the information referred to in Article 13(3)(b-f) must be described, which leaves a wide margin of interpretative freedom to the provider in specifying the content and scope of the obligation of transparency in the instructions for use.

In practice, this translates into clear legal uncertainty and a variable level of compliance by providers who, depending on their position and strength in the market, could be clearly discouraged as to the level of transparency necessary and appropriate to comply with the obligations of Article 13 of the Regulation.

2. A better articulation of the relationship between transparency, interpretability and explainability: discouraging black box models?

Although in its Recital (27), the Regulation incorporates a definition of the term "transparency", it does not, however, do the same for "interpretability" and "explainability". Recital (27) does not incorporate a definition of "transparency" as such, but rather identifies its constituent elements (traceability, explainability and communication of relevant information). While the definition of "transparency" in Recital (27) includes explainability and ignores interpretability, Article 13 seems to give greater prominence to interpretability to the detriment of explainability.

Article 13 establishes a type of transparency that is technical, internal, self-referential and limited to the relationship between the provider and the deployer. From the perspective of the provider, transparency would be aimed at regulatory compliance and at proving such compliance to the competent authorities. From the deployer's perspective, transparency would aim not only at regulatory compliance, but also at enabling the deployer to correctly interpret the output results and the proper use of the system in accordance with the provider's instructions for use.

The relationship/distinction between interpretability and explainability is unclear (e.g., in Article 13.3(d)). Article 13 seems to incentivise inherently interpretable models in (1)(3)(b)(vii)(3)(d), at least in relation to high-risk systems, rather than black box models in need of complementary explainability techniques and tools. But it is not clear, in any case, whether this was the legislator's intention, since such an approach could stifle innovation. The wording used by Article 13 seems to suggest that only a local level of explainability would be required, to the detriment of global explainability, since in its various paragraphs the interpretability requirement is limited exclusively to "system output information", thus excluding other elements of the system.

3. Indefinition of the "appropriate type and level of transparency".

The possible taxonomies and levels of transparency required to comply with Article 13 are not defined. There are expressions with a high degree of indeterminacy, such as "sufficiently transparent" or "appropriate type and degree of transparency" (Art. 13.1), which seem to modulate the transparency requirement according to criteria that are not defined in the text, which, in practice, will result in inevitable legal uncertainty for the provider and the person responsible for the deployment of the high-risk AI system.

4. A "de minimis" content of material transparency without specifying its scope.

The instructions for use provided to the person responsible for deployment shall include at least the following information:

- The identity and the contact details of the provider and, where applicable, his authorised representative (Article 13(3)(a)).

- The functional suitability of the system, including its characteristics, capabilities and limitations (Article 13(3)(b)).

- Changes to the system and its performance predetermined by the provider (Article 13(3)(c)).

- The human oversight measures envisaged, including the technical measures put in place to facilitate the interpretation of output information from the system (Article 13(3)(d)).

- The computational resources, hardware and expected lifetime of the system (Article 13(3)(e)).

- Log controls implemented (Article 13(3)(d)).

Article 13 does not specify quantitative or qualitative criteria for the content or scope of the categories of information set out. It is not yet clear how much leeway providers may have in determining the content and scope of this information. This could lead to restrictive interpretations in order to protect intellectual property rights, industrial property rights and competitiveness. In the field of standardisation, there are already approved technical standards that establish different levels of transparency depending on different taxonomies of the system level (context, system itself, model and data), and on the categories or roles of interested persons to whom the relevant information is addressed (user or system deployer, developers, auditors, control authorities, persons affected by AI systems, or the general public).

5. Standardisation and transparency of high-risk systems

The European Commission has mandated CEN and CENELEC to develop technical rules and standardisation documents to specify the content and scope of the requirements for high-risk systems set out in Section 2 of Chapter 3, including the transparency and disclosure requirement of Article 13 AIA. It is in standardisation that the actual development of rules specifying the application of the AIA will take place and where, in theory, the type and level of transparency should be specified. It is not yet clear to what extent standardisation is the appropriate instrument to incorporate technical-legal safeguards against adverse impacts on fundamental rights, and if so, how it will balance rights against innovation and the economic interests of operators in the development and commercialisation of AI.

6. The "great forgotten ones" of the AIA

Despite the communication obligations set out not only in Article 13 but in other provisions of the Regulation, the rule would in no case be truly enabling for the exercise of rights by those affected by high-risk systems in the absence of a clear framework providing individuals with clear avenues to challenge decisions taken by AI systems that affect them. Neither Article 50 (transparency obligations for providers and deployers of certain AI systems) nor Article 86 (right to an explanation of individual decision-making) ensure that the general public is provided with sufficient information to understand the risks to which they are subject and to effectively challenge individual decisions that cause adverse effects on health and safety or fundamental rights.

HUMAN OVERSIGHT OR MONITORING IN ARTICLE 14 OF THE ARTIFICIAL INTELLIGENCE ACT: A MERE MANDATORY REQUIREMENT FOR HIGH-RISK SYSTEMS?

Guillermo Lazcoz Moratinos

Centre for Biomedical Research Network (CIBERER - ISCIII)
Jiménez Díaz Foundation Health Research Institute (IIS-FJD)

I. Introduction

It was more than foreseeable that human supervision would become an essential part of the European regulation of Artificial Intelligence.

In 2019, the Commission's High Level Expert Group on AI included human supervision as one of the seven requirements for the development of trusted AI[1]. In February 2020, the Commission established human supervision as one of the mandatory requirements for high-risk AI applications in its White Paper on AI[2]. In the same year, the Parliament did the same in its proposal for a Regulation on ethical principles for the development, deployment and use of AI, robotics and related technologies[3]. Thus, in April 2021, human supervision became part of the first version of the AIA by the Commission and has remained so with a broad consensus until today.

Human oversight has been considered a fundamental ethical principle in debates about AI regulation. However, neither the terminology (as the title of this paper[4] attests - oversight? supervision? control? intervention?) nor

[1] European Commission, Directorate-General for Communication Networks, Content and Technologies, *Ethical guidelines for trusted AI*, Publications Office, 2019. Available at: https://data.europa.eu/doi/10.2759/14078

[2] European Commission, 'White Paper on Artificial Intelligence - A European approach to excellence and trust', Brussels, COM(2020) 65 final, 19 February 2020. Available at: https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligen-ce-feb2020_es.pdf, pp. 25-26.

[3] This proposal for a regulation was included in the annex to European Parliament resolution of 20 October 2020 with recommendations to the Commission on a framework for the ethical aspects of Artificial Intelligence, robotics and related technologies (2020/2012(INL)). Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_ES.html

[4] Even the English translation of "human oversight" has been confusing. In the Commission's version it was translated as "*vigilancia humana*", a term that had hardly been used in the literature or previous documents, and although it has remained so, the final version incorporates several references to "*supervisión humana*" in parallel. In this paper I opt for the term "oversight" as it is the most widely used term in the literature, and also as the most common translation of "oversight" in various documents of the EU institutions.

its definition is uniform among legal scholars or in EU policy documents. However, the more general question of whether it is necessary to include this principle in regulation seems to be out of the question. Translating this principle into a concrete rule and making it work in accordance with the objectives set by legislators (bringing human oversight down to Earth) seems a more complicated task.

This text first analyses how the Commission embodied the principle of humane supervision in the first version of the AIA and, subsequently, how it was debated by the European institutions along the legislative path to its final version. This embodiment of the principle in the regulation is crucial. In Enqvist's words, the specific design of the obligation, on whom it falls and in which contexts it surfaces, is of great importance in assessing what impact human oversight could and can have on the supervision of AI system processes.[5]

As we shall see, human oversight has incorporated only minor revisions in this legislative trajectory. On the one hand, such an uncontroversial trajectory could mean that the Commission adopted a satisfactory integration of this principle early on. In this regard, I will argue that the first version already incorporated a fairly flexible human oversight requirement for high-risk AI systems, which will be useful in different decision-making contexts. On the other hand, an uncontroversial legislative path could mean that appropriate criticisms have not been made. Or, at least, that the complexity of "bringing human oversight down to Earth" at the legislative level has not been fully observed.

Therefore, the last part of this text aims to bring this legal analysis closer to other considerations that have not been made explicit in this legislative avenue. The merits, limitations and shortcomings of human oversight in the AIA will be explored through three questions that remain open for further discussion. Namely, whether human beings can fulfil the regulatory objective of human oversight in the AIA, whether human beings are necessary *in the loop* within the decision-making processes to ensure the effective oversight required by the regulation, and whether it is human-centred beyond human oversight. In light of these reflections, it seems reasonable to assert that there is much interdisciplinary - and not just regulatory - work to be done if the AIA is not to become another policy failure of human oversight of automated systems.[6]

[5] Enqvist, L., ""Human oversight" in the EU Artificial Intelligence act: what, when and by whom?", *Law, Innovation and Technology*, vol. 15, no. 2 (2023), pp. 508-535.

[6] Other works critical of these policies include: Green, B., "The flaws of policies re-

II. Human oversight or monitoring in the Commission's April 2021 proposal

In this section I will explain how human supervision has been integrated into the AIA. I will first outline the Commission's initial proposal, then highlight the issues raised by the various institutions involved in the legislative process, and finally address the final version of the text.

Human oversight plays a prominent role among the mandatory requirements for high-risk AI systems. From the outset, the Commission set out what the regulatory objective of human oversight is (to prevent or minimise risks), what type of human oversight is required ("effective by design") and what requirements human oversight must meet to be considered effective. As we shall see, the structural points of that initial proposal have remained intact.

While this applies to all mandatory requirements, it should be noted that the AIA assigns most of the obligations relating to these requirements to the provider. That is, before placing high-risk AI systems on the market or putting them into service, providers must ensure that their high-risk AI systems comply with the mandatory requirements, and demonstrate their compliance by carrying out a quality management system, among other obligations. As we will see, since the Commission's first drafting, the obligations of deployers for the use phase have increased.

According to Article 14 of the first version of the European Commission's AIA, high-risk AI systems shall be designed and developed in such a way that they can be effectively overseen by natural persons during the period in which the AI system is in use. In other words, high-risk AI systems must by design allow for effective human supervision.

However, this mandate is only the tip of the iceberg. As Enqvist says, human oversight is not a "one-size-fits-all" requirement, and may have different orientations regarding, among others, which aspects of a system's decision-making process should be targeted, when oversight should be conducted, or who is the human being who should conduct the oversight.[7]

From this point of view, beyond this first paragraph of Article 14, there is much to disentangle from this human supervision requirement. Fortunately, and unlike, for example, Article 22 of the GDPR, which hardly provides any

quiring human oversight of government algorithms", *Computer Law & Security Review*, vol. 45 (2022), 105681; Huq, A. Z., "A Right to a Human Decision", *Virginia Law Review*, vol. *106*, no. 3 (2020), pp. 611-688.

[7] Enqvist, L., ""Human oversight" in the EU Artificial Intelligence act: what, when and by whom?", *Law, Innovation and Technology*, vol. 15, no. 2 (2023), pp. 508-535.

information on the human intervention required for automated decisions, the AIA explains in detail how human supervision must be ensured.

1. Human supervision to prevent or reduce risks

Article 14.2 of AIA's Proposal: *Human oversight shall aim at preventing or minimising the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of other requirements set out in this Chapter.*

In the explanatory memorandum, the Commission argues that human oversight throughout the lifecycle of AI systems aims to minimise the risk of algorithmic discrimination, complementing existing Union legislation on non-discrimination. Furthermore, the memorandum explains that in critical areas such as education and training, employment, important services, law enforcement and the judiciary, human oversight will also reduce erroneous or biased AI-assisted decisions, thereby facilitating the respect of other fundamental rights, in addition to non-discrimination.

In both the Ethical Guidelines for Reliable AI and the White Paper on AI we find that human supervision is mentioned as a safeguard to avoid detrimental effects of AI systems. In the scientific literature we find authors claiming that humans are crucial to avoid undue correlations and thus ensure fairness in data analysis[8], and not only to exclude discrimination, but also to reduce false positives.[9]

However, this hypothesis has been strongly contested. Among others, Huq argues that the flawed quality of a machine's decision does not imply that a human would do better[10] and that the equality problem must be addressed separately from any right to a human decision[11]. Moreover, if humans systematically fail in this task, human supervision will lead to a false sense of security[12].

[8] Favaretto, M., de Clercq, E., and Elger, B. S., "Big Data and discrimination: perils, promises and solutions. A systematic review", *Journal of Big Data*, vol. *6*, no. 1 (2019), pp. 1-27.

[9] Roig, A., "Safeguards for the right not to be subject to a decision based solely on automated processing (Article 22 GDPR)", *European Journal of Law and Technology*, vol. *8*, no. 3 (2017), pp. 1-17.

[10] Huq, A. Z., "A Right to a Human Decision", *Virginia Law Review*, vol. *106*, no. 3 (2020), pp. 611-688.

[11] Ibid.

[12] Green, B., "The flaws of policies requiring human oversight of government algorithms", *Computer Law & Security Review*, vol. 45 (2022), 105681.

And Laux reminds us that mandating human oversight is not a panacea for preventing and minimising the risks of AI.[13]

This debate, which goes to the very heart of the proposal, should not be overlooked. Especially if the Commission considers human beings as a kind of last call when all other safeguards fail, as is clear from the end of the second paragraph of this article.

2. Effective human oversight by design

While human oversight was an essential requirement in all the policy and regulatory background to this proposed Regulation, we did not find consensus on what kind of human oversight should be required. While the White Paper on AI mentioned that the appropriate type and degree of human oversight may vary from case to case[14], the European Parliament's text stated that high-risk AI systems should be subject to significant human review, assessment, intervention and control.[15]

As noted above, from its initial proposal, the AIA required that high-risk AI systems be designed and developed in such a way that they can be effectively supervised by natural persons during their use phase, including, inter alia, providing them with adequate human-machine interface (Art. 14(1) proposed AIA). Thus, the AIA requires that high-risk systems must be capable of *effective* human supervision.

Establishing this obligation from the design and development of the system means that compliance with this human oversight requirement must be ensured before the AI system is placed on the market. The Commission requires the provider to ensure that its AI systems comply with the high-risk requirements (Art. 16(a) AIA) and to establish a quality management system to document and demonstrate compliance (Art. 17(1) AIA).

Thus, the AIA establishes a governance mechanism for system design that does not necessarily determine how human oversight will be applied in the use phase of the high-risk AI system.

Let us look at Article 22 of the GDPR to illustrate the difference between the two governance mechanisms. Decisions based solely on automated processing are generally prohibited by Article 22(1), so any decision that produces a legal or similar effect on the data subject must incorporate human

[13] Laux, J., "Institutionalised distrust and human oversight of Artificial Intelligence: towards a democratic design of AI governance under the European Union AI Act", *AI & SOCIETY* (2023), pp. 1-14.

[14] European Commission, COM(2020) 65 final, p. 19.

[15] European Parliament resolution of 20 October 2020, 2020/2012(INL), Recital 10.

intervention into the data processing decision loop. Thus, Article 22(1) creates a governance mechanism based on human intervention for automated processing of personal data. And the GDPR provides that it is the controller at the use phase of an AI system that must ensure this safeguard. Moreover, the controller cannot circumvent the prohibition by artificially manufacturing human intervention and therefore controllers must ensure that any human intervention is meaningful to the decision-making process[16]. Thus, the type of monitoring-assurance required by the GDPR is meaningful human intervention for the use phase of the automated system.

The Commission, mindful of the difference between these two governance mechanisms, establishes a link between human oversight in the AIA and use-phase governance mechanisms such as Article 22 of the GDPR. Article 29(1) of the proposed AIA provides, among the obligations of those deploying high-risk AI systems[17], that they must use such systems in accordance with the instructions for use to be addressed below. However, according to the second paragraph of this Article, this obligation is *without prejudice to other user obligations under Union or national law* (such as Article 22 GDPR) *and to the user's discretion in organising its own resources and activities for the purpose of implementing the human oversight measures indicated by the provider.*

3. How to achieve an effective human supervision?

Such a governance mechanism is not only about the type or kind of human oversight that is required. Moreover, the requirement that the system can be monitored "effectively", as such, does not say much. And indeed, the same can be said of other terms used in similar governance mechanisms, such as "meaningful", or even of the differences between human "oversight" and "control", "surveillance", "intervention" or "review". In this sense, it seems that the AIA is intended to provide legal certainty for providers to comply with the requirement of effective human oversight by design.

First, it sets out what providers need to do to include human oversight in the design and development of their AI systems.

In this respect, in its third paragraph, Article 14 of the proposed Regulation indicates that the provider shall implement measures to ensure human oversight in two different ways; (a) by identifying and incorporating such

[16] Article 29 Data Protection Working Party, '*Guidelines on automated individual decisions and profiling for the purposes of Regulation 2016/679' (2019), p. 21.*

[17] "Users" in the Commission's first version, "implementers" in later versions and "deployers" in the final version.

measures, where technically feasible, into the high-risk AI system before placing it on the market or putting it into service; and/or (b) by identifying human oversight measures before placing the system on the market or putting it into service that are suitable for implementation by the deployer.

These measures will be complemented by the "instructions for use"[18] which the provider must draw up in accordance with the transparency requirement. In other words, the instructions to be received by the deployer for the use phase of the high-risk AI system must include the human oversight measures put in place by the provider (Art. 13(3)(d) AIA).

Secondly, the text provides criteria for providers to understand what constitutes "effective" monitoring.

To this end, the Commission sets out a number of capabilities that, depending on the circumstances, the human assigned to supervise the AI system must be able to perform during its use (proposed Art. 14(4) AIA). These include fully understanding the capabilities and limitations of the system, being aware of automation bias, correctly interpreting the system's output information, dismissing, overriding or reversing such information, or interrupting the system by pressing a button specifically intended for that purpose.

Therefore, to establish effective monitoring by design, the provider must implement or identify measures that allow humans – depending on the circumstances – to correctly understand and interpret the system's results, and to decide when not to use or even stop the AI system.

4. The role of the deployer in human oversight

We have already seen that the Commission focuses on providers the obligations for high-risk AI systems. This is not to say that there are no obligations on deployers in the AIA, such as those set out in section 3 on the obligations of different actors in relation to these systems.

The main obligation is that those responsible for the deployment use these systems in accordance with the instructions for use accompanying the systems. That is, to follow the instructions regarding human supervision measures. However, as noted above, following the instructions is without prejudice to other obligations of the deployer under Union or Member State law and to the user's discretion in organising its own resources and activities. In order to comply with this provision, the application of Article 22 GDPR

[18] Art. 3(15) AIA: "Instructions for use" means *the information provided by the provider to inform the deployer of, in particular, an AI system's intended purpose and proper use.*

comes into play, but also Article 11.1 of Directive (EU) 2016/680[19] or Article 7.6 of Directive (EU) 2016/681[20], among others. This will depend on the context of use of the high-risk system.

In addition, another obligation that may affect the way in which human oversight is carried out is that deployers shall ensure that the input data is relevant in view of the intended purpose of the high-risk AI system (Art. 29(3) AIA). In AI-based decision-making contexts, human operators may be assigned the role of reviewing the input data for such decisions. Especially where such data may be of sensitive categories. In addition, such checks could be put in place both before decisions are taken (*ex ante*), and afterwards to correct erroneous decisions (*ex post*).

Following what De Hert and I have argued in another article[21], Article 29(6) of the AIA establishes a link between the provider's obligations under this Regulation and the controller's obligations under the GDPR (where they are themselves data controllers under the GDPR). Deployers of high-risk AI systems will make use of the information provided by AI system providers under the transparency requirement *to fulfil their obligation to carry out a data protection impact assessment under Article 35 of Regulation (EU) 2016/679 (…), where applicable*. As noted above, this information includes human oversight measures. In other words, deploying controllers (data controllers) will receive technical and organisational information (from AI system providers) about the AI systems they acquire and are obliged to make use of this information - to comply with Article 35 GDPR - which enables natural persons (in the controller's organisation) who are assigned human oversight in Article 22 GDPR to understand the capabilities and limitations of the system. We consider promising this link between the AIA and the GDPR.

Finally, for systems intended to be used for "real-time" or "delayed" remote biometric identification of natural persons, the Commission established

[19] Directive (EU) 2016/680. Article 11(1): *Member States shall provide for the prohibition of decisions based solely on automated processing, including profiling, which produce adverse legal effects on the data subject or significantly affect him or her, unless authorised by Union or Member State law to which the controller is subject and which provides for appropriate measures to safeguard the rights and freedoms of the data subject, at least the right to obtain human intervention by the controller.*

[20] Directive (EU) 2016/681. Article 7(6): *The competent authorities shall not take any decision that produces an adverse legal effect on a person or significantly affects a person solely by reason of the automated processing of PNR data. Such decisions shall not be based on a person's race or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, health or sexual life or orientation.*

[21] Lazcoz, G., and de Hert, P., "Humans in the GDPR and AIA governance of automated and algorithmic systems. Essential pre-requisites against abdicating responsibilities", *Computer Law & Security Review*, vol. *50* (2023), 105833.

what appears to be an intermediate obligation between providers and deployers. According to this paragraph, providers' measures for these systems must ensure that the deployer does not act or make any decision on identification generated by the system unless it has been verified and confirmed by a minimum of two natural persons.

III. The trajectory of human oversight in the ordinary legislative procedure

Although other amendments and proposals can be found in this legislative process, this section focuses on the two key issues that have been discussed on human oversight in the AIA.

On the one hand, the right to human intervention has been discussed. As the Commission focused on establishing obligations for providers that are to place high-risk systems on the market, it did not declare a right as such to human supervision at the stage of use of AI systems. Thus, the initial version obliges providers to market AI systems that can be effectively supervised by natural persons. However, the type of supervision should be determined on the basis of the applicable regulation (e.g., Article 22 of the GDPR) and at the discretion of the deployer itself. Contrary to this regulatory approach, there have been calls throughout the legislative process for the inclusion in the AIA of a right of human intervention for high-risk AI-based decisions.

On the other hand, the role of humans in overseeing high-risk systems has also been debated. The involvement of humans in AI-based decision-making will not magically remedy the harmful effects of these automated systems. In this sense, Matsumi and Solove humorously define the way in which this normative role is usually established for people: *For human involvement to be the answer, the law must set forth exactly how humans would ameliorate the problems with algorithmic predictions in particular cases. Instead, the law just points to a human and says: "Hey, there's a human, so all is fine" even though it remains unclear what the human is to do*[22]. This has been one of the main problems with the human-based governance mechanisms referred to earlier in this text. While the Commission spells out the requirements of the system to allow for effective human oversight, it says nothing about the human beings who are entrusted with this task.

[22] Matsumi, H., and Solove, D. J., "The Prediction Society: Algorithms and the Problems of Forecasting the Future", *GWU Legal Studies Research Paper*, vol. *58* (2023), pp. 1-64.

1. Voices calling for the right to human intervention (and other safeguards) for decision-making based on high-risk systems

For many authors, the Achilles' heel of the Commission's proposal lay in the (non-existent) rights of individuals subject to AI-based decisions. And not only rights in relation to the use of the systems, but also mechanisms that would allow users affected by AI to hold the various actors involved in the lifecycle of the systems accountable. And this was the direction in which these early criticisms of the AIA were voiced, for example, in the words of Veale and Borgesius: "*As only those with obligations under the Draft AI Act can challenge regulators" decisions, rather than those whose fundamental rights deployed AI systems affect, the Draft AI Act lacks a bottom-up force to hold regulators to account for weak enforcement.*[23]

Regarding human oversight, there have been calls for the inclusion in the AIA of different rights for the use phase based on this mandatory requirement. In particular, the right to human intervention, or the right to a *human in the loop*, in decision-making processes with high-risk AI systems. However, there have also been references to the right to an explanation or the right to challenge automated decisions, which would also be mediated by human operators adopting a reviewer role.

Firstly, the European Committee of the Regions called for a right to human involvement in any decision taken by high-risk AI systems. In the terms used by the Committee, such a decision *will be subject to human intervention and based on a rigorous decision-making process. Human contact with these decisions must be guaranteed.*[24]

The opinion of the European Economic and Social Committee (EESC) is along the same lines, albeit with more elaborate arguments[25]. The Committee wonders whether we are ready for AI to substantially take over the role of human decision-making, even in critical processes. Among the critical areas where these decisions have significant moral and legal implications or social impact, the EESC mentions the judiciary, law enforcement, social services,

[23] Veale, M., & Zuiderveen Borgesius, F., "Demystifying the Draft EU Artificial Intelligence Act - Analysing the good, the bad, and the unclear elements of the proposed approach", *Computer Law Review International*, vol. *22*, no. 4 (2021), pp. 97-112.

[24] Opinion of the European Committee of the Regions - A European approach to Artificial Intelligence - Artificial Intelligence Act (revised opinion). COR 2021/02682.

[25] Opinion of the European Economic and Social Committee on the Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislation (COM(2021) 206 final - 2021/106 (COD)) - EESC 2021/02482.

healthcare, housing, financial services, labour relations and education. Hence the EESC recommends that in these areas decisions *remain with the people*.

As regards the amendments adopted by the European Parliament, recital 58a of the proposal highlights the key role of the deployers in ensuring the protection of fundamental rights. As deployers are best placed to understand how the AI system will be used in a specific context, they should identify the appropriate governance structures for that context. In Parliament's first reading, such appropriate governance structures include: *complaints handling procedures and redress procedures, as choices in governance structures can be decisive in mitigating risks to fundamental rights in specific use cases.*[26]

Again, a policy conflict arises between establishing in the AIA a universal set of governance mechanisms for human oversight at the use phase or, on the contrary, allowing more discretion depending on the specific context of use (which may also be limited by the applicable regulations in each context).

However, Parliament also wanted to include in the AIA a right to an explanation of individual decision-making. According to Article *68c* of this version, any affected person who is the subject of a decision based on the results of a high-risk AI system that produces legal or significant effects *shall have the right to request from the deployer a clear and meaningful explanation (…) of the role of the AI system in the decision-making procedure, the main parameters of the decision taken and the corresponding input data*[27]. At first glance, this right seems linked to a logic of human oversight (it forms a transparency requirement mediated by a kind of human oversight of the decision-making process) whereby the decision of the AI system is brought closer to the understanding of the individual concerned.

Finally, chronologically speaking, Opinion 44/2023 of the European Data Protection Supervisor (EDPS) also addressed this issue and called for the inclusion of a right to obtain human intervention of the (end) user of the AI system in relation to decision making that affects him/her and to challenge the outcome of the decision making, as well as a right to receive explanations from the controller of the AI system about decision making that significantly affects him/her[28]. The EDPS considers that such rights would not affect, but would complement the rights established by Article 22 of the

[26] Amendment 92.

[27] Amendment 630.

[28] European Data Protection Supervisor, Opinion 44/2023 on the Proposal for Artificial Intelligence Act in the light of legislative developments, 23 October 2023, Brussels, p. 25. Available at: https://www.edps.europa.eu/data-protection/our-work/publications/opinions/2023-10-23-edps-opinion-442023-artificial-intelligence-act-light-legislative-developments_en.

GDPR, and other rights established by the applicable law in each context of use, such as consumer credit, insurance services, employment, etc.[29]

2. Humans, what humans?

Who (and under what conditions) is in charge of overseeing an AI system was also the subject of debate in this legislative route. Koulu explains that EU policy documents raised high expectations about human oversight to safeguard people's autonomy, somewhat mystifying human capabilities as a last line of defence against this flood of external intelligence[30]. Thus, in these EU policy documents we find no discussion of the implications of human supervision, whether and how human supervisors are capable of performing their supervisory tasks, nor what the criteria for human intervention would be or whether a supervisor should possess any particular expertise.[31]

If we look at the first version of the AIA, we can stick to Koulu's view on previous policy documents. Only recital 48 of the AIA proposal refers to the capacity of human beings to carry out the task of monitoring: *In particular, where appropriate, such measures should guaranteee (…) and that the natural persons to whom human oversight has been assigned have the necessary competence, training and authority to carry out that role.* These words, however, do not translate into any obligation for those responsible for deployment in the articles of the Commission's proposal.

In their Joint Opinion of 2021, the CEPD-SEDP advocated for a true human centrality which should be supported by highly qualified human supervision[32]. Among the various safeguards necessary to ensure that the rights of data subjects are respected and guaranteed and to avoid negative effects on individuals, in particular on the production of biased decisions, the CEPD-SEDP emphasised a qualified human supervision in such decision-making processes. In this context, they consider that competent authorities should

[29] Ibid, pp. 17-18.

[30] Koulu, R., "Human control over automation: EU policy and AI ethics", *Journal of Legal Studies*, vol. 1 (2020), pp. 9-46.

[31] Ibid.

[32] CEPD-SEPD, Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act), 18 June 2021, Brussels, p. 6. Available at: https://www.edpb.europa.eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-52021-proposal_es

also be able to propose guidelines to assess biases in AI systems and assist the exercise of human oversight.[33]

In terms of labour relations, and given that monitoring will be carried out by a worker or a group of workers, the European Economic and Social Committee stressed that these workers should be trained on how to perform this task: *Furthermore, as it is expected that these workers may ignore the outcome of the AI system or even not use it, measures should be put in place to avoid fear of reprisals (such as demotion or dismissal) in case such a decision is taken* (4.18)[34]. The Committee therefore calls for specific measures for the decision-making process when humans feel that they should disobey the high-risk AI system. Granting this authority to workers may be a necessary condition to avoid automation bias.

The European Economic and Social Committee also notes that the Commission's AIA proposal lacks a forward-looking vision that highlights the potential of AI to augment, rather than replace, human decision-making.

When the text reaches the European Parliament's first reading, these ideas are translated into the introduction of new obligations for deployers. In this version of the AIA, among the obligations of both providers and deployers in Article 16, it is necessary to ensure that *natural persons assigned to human supervision of high-risk AI systems are, in particular, aware of the risk of automation or confirmation bias;*[35]. This provision therefore includes specific measures on the part of the provider to enable this control of automation and confirmation bias and, on the part of the deployer, measures to ensure that the specific persons assigned to oversee are aware of these risks.

In addition, Parliament included new obligations for deployers in Article 29, according to which they shall (i) implement *human oversight in accordance with the requirements set out in this Regulation*; ensure (ii) *that natural persons responsible for the human oversight of high-risk AI systems are competent, appropriately qualified and trained, and have the necessary resources (…); and ensure that (iii) relevant and appropriate robustness and cybersecurity measures are regularly monitored for effectiveness and regularly adjusted or updated (…).)*[36]. This last paragraph is therefore promising, as it includes the notions of competence, qualification, training, and resources needed for the humans who are supposed to effectively oversee high-risk AI systems.

[33] Ibid, p. 17.

[34] European Economic and Social Committee, COM(2021)0206 - C9-0146/2021 - 2021/0106(COD).

[35] Amendment 334.

[36] Amendment 401.

IV. The final version of human oversight in the AIA at a glance

The final version of the AIA as a result of this legislative process yields the following key features of human supervision.

- High-risk AI systems that can be effectively overseen by natural persons by design (Art. 14(1) AIA).

- Human oversight to prevent and reduce risks, particularly when other safeguards for high-risk systems are not effective (Art. 14(2) AIA).

- Measures proportionate to the risks, level of autonomy and context of use of the high-risk system that can be technically integrated by the provider and/or defined by the provider to be implemented by the deployer (Art. 14(3) AIA).

- Measures aimed at ensuring that the person in charge of monitoring has a proper understanding of the capabilities and limitations of the system, is aware of the automation bias, correctly interprets the information output from the system, can decide in a specific situation not to use the system and/or even stop the system if necessary (Art. 14(4) AIA).

- Deployers implementing appropriate technical and organisational measures to ensure that systems are monitored in accordance with the instructions and measures implemented by the provider (Art. 26(1) AIA).

- Human oversight entrusted to natural persons having the necessary competence, training and authority, as well as the necessary support (Art. 26(2) AIA).

- Human oversight by design that does not affect compliance with national and EU regulatory obligations regarding decision making at the use phase, or the freedom to organise the resources and activities of the deployer (Art. 26(3) AIA).

- Remote biometric identification systems limited in decision-making to being verified and confirmed separately by at least two natural persons (Art. 14(5) AIA).

V. Some reflections on what we can expect from human oversight in the AI Act

The EU is not the only government institution seeking to regulate AI, but it is one of the most advanced in this task. The symbolic potential of this regulation – deliberately sought by the European legislator – is also no secret. In the case of human supervision, Article 14 of the AIA is likely to form the basis of the first general provision on the subject, and is therefore likely to attract a lot of attention and serve as a test of general supervision requirements.[37]

It is therefore necessary to consider what we can expect from this novel provision that will set the tone for other regulations. In this section, I conclude by exploring some open questions about the merits, limitations, and shortcomings of human oversight in the AIA.

The first concerns the demands that such governance mechanisms place on humans. Some authors have been highly critical of other regulations that require human oversight of automated systems because they have been ineffective, not least because of the inability of humans to meet the regulatory objectives they set out.

The second is about the type of human oversight that AIA requires, and whether this type of human oversight can work in the real world. Designing what type of human oversight is appropriate for each decision-making context is not an easy task[38]. The regulation must therefore strike a difficult balance between ensuring sufficient flexibility to design human oversight for each context and imposing common minimum standards of oversight.

The last question seeks to explore the human rationality of human supervision itself. The call for human oversight has been linked to the development of the concept of human-centred AI. Indeed, human oversight is said to be a procedural and reactive approach to "human-centred" AI[39]. But what are the implications of the relationship between the two concepts and is human oversight in human-centred AIA?

1. Can human beings fulfil the normative purpose of human oversight in the Regulation?

As mentioned above, human oversight in the AIA aims to prevent or minimise risks to health, safety or fundamental rights.

However, this approach is controversial. Huq argues that the flawed quality of a decision made by a machine does not guarantee that a human being

[37] Enqvist, L., "“Human oversight” in the EU Artificial Intelligence act: what, when and by whom?", *Law, Innovation and Technology*, vol. 15, no. 2 (2023), pp. 508-535.

[38] Yurrita, M., Draws, T., Balayn, A., Murray-Rust, D., Tintarev, N., and Bozzon, A., "Disentangling Fairness Perceptions in Algorithmic Decision-Making: The Effects of Explanations, Human Oversight, and Contestability", in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023).

[39] Enqvist, L., "“Human oversight” in the EU Artificial Intelligence act: what, when and by whom?", *Law, Innovation and Technology*, vol. 15, no. 2 (2023), pp. 508-535.

would do better[40], and that the problem of equality and non-discrimination must be addressed separately from any right to a human decision[41]. Similarly, Green argues that human oversight policies are not supported by empirical evidence and are therefore unlikely to protect against the harms of algorithmic decision-making.[42]

On the other hand, the EU's legal-political documents created high expectations on human supervision to safeguard human autonomy in the development and use of AI, which is not a good idea according to Koulu[43]. The technological focus of these documents ends up assigning subjectivity to AI while mystifying human capabilities. In this way, human supervisors - human agents involved in AI decision-making processes - are presented as the last line of defence against AI, while AI is anthropomorphised into an autonomous agent that could be malicious towards humans.[44]

In my opinion, these are two sides of the same coin.

I agree with Green and other authors who have analysed human oversight in different governance mechanisms, their real-world functioning is far from optimal and, at the very least, we can say that they are not fulfilling the normative objectives for which they are designed. This is not to say that humans cannot (should not) play a decisive role in decision-making with high-risk AI systems. In fact, it does not seem that the humans controlling these decision-making processes have done so badly so far. And we have equipped ourselves with "classical" legal mechanisms for cases where such human control of decision-making is inadequate or fails. With AI, this paradigm of human decision-making seems to change. However, does this mean that humans cannot contribute to improving AI-guided decision-making in the socio-cultural contexts in which it is applied?

At this point, what we seem to need are evidence-based governance mechanisms for human oversight. That is, not just stating in the abstract that

[40] Huq, A. Z., "A Right to a Human Decision", Virginia Law Review, vol. 106, no. 3 (2020), pp. 611-688. However, the basis for the claim that machines outperform humans needs to be assessed. Many studies are based on comparisons between AI and individual performance that have little or no practical relevance, see. Cabitza, F., "*Many say that AI can outperform human doctors. Is it true?*", LinkedIn (2018). Available at: https://www.linkedin.com/pulse/many-say-ai-can-outperform-human-doctors-true-federico-cabitza/

[41] Ibid.

[42] Green, B., "The flaws of policies requiring human oversight of government algorithms", *Computer Law & Security Review*, vol. 45 (2022), 105681.

[43] Koulu, R., "Human control over automation: EU policy and AI ethics", *Journal of Legal Studies*, vol. 1 (2020), pp. 9-46.

[44] Ibid.

humans reduce AI risks, but providing mechanisms that do so effectively. This would mean requiring via regulation, throughout the entire AI system cycle, that it is designed and implemented in such a way that humans are able to demonstrably reduce the risks involved in the decision-making process. Does the AIA provide evidence-based human oversight for high-risk AI systems?

2. Is it necessary, under the Regulation, to have humans in the loop of high-risk AI decision-making to ensure the required effective human oversight?

In response to fears of automation, human oversight at the regulatory level has historically been associated with a humankind maintaining the final say over an automated system. This ensures that the results provided by an automated system are not the sole reason for decision-making, as the human operator can change the system's criteria until the final decision is made[45]. In this way, "*human in the loop*" has become a standard regulatory solution to solve the problems of transparency, bias, legal certainty and systemic risks related to automation.[46]

However, the complexity of hybrid human-machine decision-making processes is increasing due to technological and social progress. For example, looking at the use case scenarios presented by Enarsson, Enqvist and Naarttijärvi, we can conclude that hybrid decisions are an amalgam of legal, social, technical and organisational issues[47]. Hence, it is difficult to find a single legal solution, such as giving the final say to the people involved in such processes.

The fact is that, although modern AI systems considerably reduce the importance of humans in decision-making processes, humans are still involved in them in countless ways. Matsumi and Solove put it this way: *There are humans behind every algorithmic prediction, much like the Wizard of Oz was a man operating a machine*[48]. Thus, the key is to determine, for different contexts and among all the humans involved, who is there to ensure the requirement of human supervision and what is expected of them. Their role does not have

[45] Wagner, B., "Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems", *Policy & Internet*, vol. *11*, no. 1 (2019), pp. 104-122.

[46] Enarsson, T., Enqvist, L., and Naarttijärvi, M., "Approaching the human in the loop - legal perspectives on hybrid human/algorithmic decision-making in three contexts", *Information & Communications Technology Law*, vol. *31*, no. 1 (2022), pp. 123-153.

[47] Ibid.

[48] Matsumi, H., and Solove, D. J., "The Prediction Society: Algorithms and the Problems of Forecasting the Future", *GWU Legal Studies Research Paper*, vol. *58* (2023), pp. 1-64.

to be the final say in all decisions. As Fosch-Villaronga and Malgieri point out, in certain contexts direct human intervention may be ineffective or even detrimental.[49]

According to the AIA, providers will have to design AI systems that allow humans to correctly interpret their results or decide not to use them, among others. However, the AIA does not impose that the decision-making process using the AI system has to occur one way or the other.

If we consider the AIA as a focused starting point on how to design a system so that it can be monitored effectively and to establish smooth communication between providers and deployers, this is good news. Given that the AIA is a general standard for many types of AI systems, one of its strengths is that it does not limit the type of human oversight that should be applied in the use phase.

The bad news is that we have to assume that the existing rules in the different contexts of application are sufficient to provide legal certainty - and we have already concluded that this is not the case. Or that controllers have good resources and guidance to implement this requirement in each context - probably not the case either. The risk is that the AIA will fall into the long line of failed regulatory attempts to address the complex human-machine interaction.[50]

In fact, the exception within the AIA to the leeway it gives for the use phase is in relation to remote biometric identification systems. Recital 73 of the AIA sets out the need to establish a "strengthened" mechanism for these systems[51], whereby the deployer may not act or take any decision on the basis of the identification generated by the system unless it has been verified and confirmed separately by at least two natural persons. It is striking that the AIA considers this separate double verification to be a "strengthened" mechanism - is there reason to believe that the second verifier will not incur the same bias or error as the first supervisor, and can this "separate" supervision mechanism be considered more strengthened than a "joint" decision-making process involving two human supervisors?

[49] Fosch-Villaronga, E., and Malgieri, G., "Queering the ethics of AI", in Handbook on the Ethics of Artificial Intelligence. Edward Elgar Publishing (2024), forthcoming.

[50] Beck, J., and Burri, T., "From "Human Control" in International Law to "Human Oversight" in the New EU Act on Artificial Intelligence". In D. Amoroso & F. Santoni de Sio (Eds.), *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems*. Elgar (2023).

[51] Among the remote biometric identification systems considered high-risk by Annex III, Art. 14(5) AIA provides for a derogation from the application of this requirement in the areas of law enforcement, migration, border control or asylum, in case an EU or national regulation considers this requirement to be disproportionate.

So, with the advent of the AIA, we still lack evidence-based human oversight governance mechanisms for the use phase of high-risk AI systems - which legislators will take the lead on this? In addition, we also need resources and guidance for those responsible for deployment to implement human oversight - which institutions will do this?

3. Beyond human supervision, do we have a human-centred Regulation?

The concept of human-centred AI has been at the heart of policy discussions on these technologies. In the White Paper on AI, the Commission strongly supported a human-centred approach as a key element for the future regulatory framework. Furthermore, it stated that the goal of trustworthy, ethical and human-centred AI can only be achieved by ensuring adequate human oversight of high-risk AI applications.[52]

Although this concept did not make it into the Commission's AIA Proposal, it has finally found a prominent place in the final version. Thus, in its first article, it states that the objective of this Regulation is, among others, to promote the adoption of human-centred and reliable Artificial Intelligence (AI). As a definition of this concept within the AIA itself, we only find in Recital 6 that AI must be a tool for people and have the ultimate goal of increasing human well-being.

If we turn to the scientific literature, Enqvist acutely explains that human supervision plays a procedural and reactive role in the "human-centred" concept of AI. While the goal of human-centred AI applications is to proactively - by design - meet human needs and preferences in different contexts, human monitoring measures seek to reactively address the risks, biases and harms of AI systems.[53]

Thus, we can see how this procedural and reactive approach has been brought to AIA through human supervision as a mandatory requirement for high-risk systems. Of course, there is no place here to debate whether the regulation - as a whole - embodies a human-centred AI policy. However, it does seem appropriate to consider whether human oversight itself, as a mandatory requirement in the AIA, is human-centred.

This approach is highly relevant because human supervisors are in fact very likely to find themselves in a vulnerable situation.

On the one hand, evidence shows how people will see their skills and abil-

[52] European Commission, COM(2020) 65 final, p. 21.

[53] Enqvist, L., ""Human oversight" in the EU Artificial Intelligence act: what, when and by whom?", *Law, Innovation and Technology*, vol. 15, no. 2 (2023), pp. 508-535.

ities diminished by being the supervisors of AI. For example, with the use of automated systems, humans do not develop the skills and knowledge that are normally acquired through experience - the "deskilling" effect -[54]. Moreover, the pervasive influence that technology may have on the user will inevitably be accompanied by side effects, such as complacency and automation biases[55]. Additionally, biased algorithmic recommendations could negatively influence human behaviour in the long run, i.e., when the automated system has withdrawn from the decision-making process.[56]

On the other hand, there are also legal side effects that will affect human supervisors. Green argues that human oversight provisions shift the responsibility for AI harms from the heads of the institutions that deploy the systems (and determine the structure of the systems) to the frontline human operators (who are relatively powerless in this regard). Human oversight policies therefore create a loophole that allows companies to adopt flawed AI systems and avoid taking responsibility for the resulting harms.[57]

We need to think about what human values we want to preserve in our interaction with technology. In terms of human supervision, this means thinking of the supervising individual not only as an intermediary between the AI system, the person about whom decisions are made, and the risks of this interaction. In other words, human supervisors should be treated as an end in themselves. Legislators, but also providers and those responsible for deploying the systems, need to reflect on how the exercise of this function affects the people entrusted with this supervision in terms of their work performance, skills, well-being, and so on. Most importantly, their rights as workers and as human beings.

Respect for human dignity and personal autonomy must include judges, doctors, public officials, content moderators, drivers, or anyone else who must oversee high-risk AI systems.

[54] Sutton, S. G., Arnold, V., and Holt, M., "How Much Automation Is Too Much? Keeping the Human Relevant in Knowledge Work", *Journal of Emerging Technologies in Accounting*, vol. *15*, no. 2 (2018), pp. 15-25.

[55] Cabitza, F., Campagner, A., Angius, R., Natali, C., and Reverberi, C., "AI Shall Have No Dominance: On How to Measure Technology Dominance in AI-Supported Human Decision-Making". In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023).

[56] Vicente, L., and Matute, H., "Humans inherit Artificial Intelligence biases", *Scientific Reports*, vol. *13*, no. 1 (2023), 15737.

[57] Green, B., "The flaws of policies requiring human oversight of government algorithms", *Computer Law & Security Review*, vol. 45 (2022), 105681.

VI. Conclusions

In the introduction to this paper I highlighted how foreseeable it is that human oversight will become an essential part of European regulation of Artificial Intelligence. As we have seen, it does so as a mandatory requirement for high-risk AI systems.

Its legislative process has been uncontroversial and, with the exception of some of the details noted above, has maintained the essence of the Commission's initial proposal until its final adoption. The fundamentals of human oversight or supervision as a mandatory requirement oblige providers to establish measures from the design stage that allow systems to be effectively supervised with the aim of reducing risks.

In the reflections in the third section, I wanted to highlight some of the difficulties that this governance model established by the AIA will face. Among these, I would like to stress, on the one hand, the difficulty for the human beings entrusted with oversight to be able to reduce the risks of these systems without evidence-based governance models. On the other hand, the difficulty of integrating, in such diverse contexts of use, effective human oversight from design when regulation at these stages has proved unsuccessful and providers and deployers have few resources to rely on.

Ultimately, while I am optimistic about the model of human oversight established by the AIA, I believe that it will require efforts by all actors involved in the development, use, and governance of AI systems for that optimism to materialise into people being able to reduce the risks of these high-risk systems effectively in a variety of contexts.

ACCURACY AND ROBUSTNESS OF HIGH-RISK ARTIFICIAL INTELLIGENCE SYSTEMS IN ARTICLE 15 OF THE ARTIFICIAL INTELLIGENCE ACT

Ana Aba Catoira

Senior Lecturer in Constitutional Law. University of A Coruña[1]

I. Introduction to accuracy and robustness in high-risk AI systems

It is indisputable that AI has great transformative potential and that it also poses inherent risks in its use. On the other hand, it is undeniable that AI is not built in a context free from discriminatory or inequitable practices[2]. That said, AI cannot be understood solely and exclusively as a technique because it has a social and ethical dimension that presupposes that a reliable and responsible AI has to be more than just a good system, i.e., a system that does what it is intended to do[3]. In this context, transparency, quality of data sets, as well as testing, evaluation, validation and verification[4] are essential elements.

In this regard, the National Institute of Standards and Technologies has identified the following technical and socio-technical characteristics necessary to cultivate trust in AI systems: accuracy, explainability and interpretability, privacy, reliability, robustness, security and security resilience, and that harmful biases are mitigated or controlled.[5]

[1] This work is carried out in the framework of the Project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by MCIN/AEI/10.13039/501100011033/ ERDF, EU.

[2] In this sense, the deepening digital divide(s) indicate that a large part of the world's population does not participate in either the design or the development of technology and, more specifically, in AI. This lack of opportunities is more evident in women and other historically discriminated social groups, which puts the focus on this idea, that is, everything related to technological development has a relevant ethical dimension, as the social consequences of the absence of women and other social groups, or at least their lesser participation, in AI developments will be evident, Aba Catoira, Ana, "Discrimination through public data without a gender perspective and digital discrimination", *(Des)igualdad y violencia de género: el nudo gordiano de la sociedad globalizada*, Ramos Hernández, Pablo (coord.), Aranzadi, Madrid, (2020), pp. 29-51.

[3] Aba Catoira, Ana, "La garantía de los derechos como respuesta frente a los retos tecnológicos", *Derecho Público de la inteligencia artificial,* Balaguer Callejón, Francisco and Cotino Hueso, Lorenzo (dirs.), Fundación Manuel Giménez Abad, Colección Obras Colectivas 27, Zaragoza, (2022), pp. 57-84.

[4] National Institute of Standards and Technology (NIST), *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.*

[5] See *Transparencia y explicabilidad de la inteligencia artificial*, Cotino Hueso, Lorenzo/Castella-

A high-risk AI system is high-risk precisely because of the potential risks that its use poses to the health, safety and fundamental rights of individuals[6]. In this sense, these systems must be prepared to minimise and prevent these risks or, in other words, harmful and undesirable behaviours, as well as being able to detect them when their operation takes place outside the domain of entry and execution established by their intended purpose. It must also be designed and implemented to avoid making wrong decisions or generating wrong output information. This is to avoid negative consequences for individuals.[7]

In this study, we will address the regulation of accuracy and robustness requirements for high-impact or high-risk Artificial Intelligence systems in AIA by analysing the legislative proposal presented by the European Parliament and the European Commission in April 2021[8], as well as the Council's

nos Claramunt, Jorge (eds), Tirant lo Blanch, Valencia, (2022); Ortiz de Zárate Alcarazo, Luis, "Explicabilidad (de la inteligencia artificial)", *Eunomía. Revista en Cultura de la Legalidad,* n.º 22, (2022), pp. 328-344.

[6] These issues have been addressed in previous works, Aba Catoira, Ana, "Derechos de igualdad, personas con discapacidad y mayores en el entorno digital (VIII, XI y XII)", *La Carta de Derechos Digitales*, Cotino Hueso, Lorenzo (coord.), Tirant lo Blanch, Valencia, (2022), pp. 123-154. 123-154; "Las garantías de los derechos en el espacio digital: La constitucionalización de lo digital", *Inteligencia artificial y democracia: garantías, límites constitucionales y perspectiva ética ante la transformación digital,* Castellanos Claramunt, Jorge (coord.), Tirant lo Blanch, Valencia, (2023), pp. 87-114; "La era de la ciudadanía conectada: digitalización y retos del futuro desde una perspectiva de género", *Un estudio sobre el Estado autonómico: propuestas de mejora para el tercer decenio del Siglo XXI*, Castellanos Claramunt, Jorge (coord.), Tirant lo Blanch, Valencia, (2023), pp. 165-190.

[7] The Artificial Intelligence Act establishes three categories of AI systems according to risk: prohibited practices (Chapter II), high-risk systems (Chapter III) and general-purpose models (Chapter IV). They are classified by applying a risk management system, clearly influenced by the General Data Protection Regulation together with other elements specific to the private sector, and, depending on that risk, they are subject, as will be seen, to different requirements and obligations proportional to the risk that their use poses to health, security and fundamental rights.

The classification of the systems and, precisely, the specific regulation of high-risk Artificial Intelligence systems constitute for Gamero "the keystone of the whole regulation", Gamero Casado, Eduardo, "El enfoque europeo de Inteligencia Artificial", *Revista de Derecho Administrativo,* CDA, n.º 20, (2021)*,* pp. 268-289*,* specifically p. 277.

For high-risk AI systems, the European standard sets out the classification criteria in Article 6.

[8] COM (2021) 206 final 2021/0106 (COD) of 21 April 2021. Disponible en: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_1&format=PDF

general approach on the proposal adopted by the Transport, Telecommunications and Energy Council at its 3917th meeting on 6 December 2022, which sets out the Council's provisional position on this proposal and formed the basis for the preparations for the negotiations with the European Parliament. No 3917 held on 6 December 2022, which sets out the Council's provisional position on this proposal and formed the basis for the preparations for the negotiations with the European Parliament[9]. This general approach introduces significant developments, for example a new definition of AI crucial for determining the scope of application of AI regulation or the reference to general purpose AI systems (so far absent and a cause of great concern in the EU). On 11 May 2023 the Internal Market Committee and the Civil Liberties Committee adopted a draft negotiating mandate on these rules which allowed a new version to be adopted incorporating the amendments adopted by the European Parliament on 14 June 2023[10]. The long-awaited text was adopted on 13 February 2024.

In this study, we will focus on analysing part of the content of Article 15 (Section 2, Chapter III), specifically the regulation of the accuracy and robustness requirements for high-risk Artificial Intelligence systems[11], which

The Annexes to the proposal, available at: https://eur-lex.europa.eu/resource.html?uri=-cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0008.02/DOC_2&format=PDF

[9] COUNCIL SECRETARIAT GENERAL, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Regulation) and amending certain legislative acts of the Union - General Approach (6 December 2022). Available at: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CONSIL:ST_15698_2022_INIT-

[10] Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_ES-.html

[11] These requirements are rooted in the *Ethical Guidelines for Trusted AI* developed by the Independent High Level Expert Group on Artificial Intelligence, established by the European Commission in June 2018. Trusted AI must respect all applicable laws and regulations, must respect ethical principles and values, and must be robust both from a technical perspective and taking into account its social environment. In addition, the guidelines present a set of 7 key requirements that AI systems must meet to be considered trustworthy: (i) human action and oversight, (ii) technical robustness and security, (iii) privacy and data management, (iv) transparency, (v) diversity, non-discrimination and equity, (vi) environmental and social well-being, and (vii) accountability. It is within the transparency requirement that the need for AI models to be explainable is integrated, and a set of criteria for assessing the extent to which a model meets these requirements is proposed. EUROPEAN UNION, *Ethical guidelines for trustworthy AI*. European Commission, Brussels, 2019. In any case, the relevance of the principle of explainability should be emphasised insofar as it underpins all the others. Indeed, it seems difficult to think that AI can be fair if the explainability of the system is not guaranteed, as well as taking into account that it is essential, not only from an ethical perspective, but also

are so classified on the basis of their functionality, purpose and use, in accordance with current legislation on product safety and use.

That said, the importance of the various issues that occupy the analysis of this paper is abundantly clear. Indeed, AI is a set of transformative technologies that are evolving unstoppably. Personalised recommendation systems, virtual assistants, autonomous cars or infection prediction, among many other applications, demonstrate on a daily basis their great capacity to improve efficiency in many areas of life, but also their potential to generate risks and threats that, if they materialise, would cause damage, even irreparable damage to public interests and the rights of individuals (Recitals 3 and 4)[12]. This is largely because intelligent systems continuously learn and adapt as they process large amounts of data, so that their behaviour can vary and evolve over time. Furthermore, algorithms can provide decisions based on highly complex mathematical models that are extremely difficult to analyse and understand.

The European initiative to approve the first comprehensive law on AI responds to this need in order to provide an adequate framework for the protection of the people who should be at the centre of technological development and who, in this sense, oblige technology to meet high standards of trust, security and freedom[13]. The Regulation aims to provide such a legal framework,

insofar as it is essential to be able to exercise control and demand accountability in the technological field that is the basis of ethical and reliable AI.37), which states that transparency and explainability of AI systems are often fundamental preconditions for ensuring respect, protection and promotion of human rights, fundamental freedoms and ethical principles. In addition, these principles allow to know why certain decisions are made and the processes by which these decisions are taken, so that such information provides guarantees for claims or complaints against them.

[12] Royal Decree 817/2023 of 8 November, which establishes sandboxes for testing compliance with the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence, also notes https://www.boe.es/eli/es/rd/2023/11/08/817; "Artificial Intelligence is a disruptive technology with a high capacity to impact the economy and society. At the economic level, and together with other digital technologies, it has a high potential for increasing productivity, opening up new lines of business, developing new products or services - based, for example, on personalisation, optimisation of industrial processes or value chains - improving the ease of performing everyday tasks, automating certain routine tasks and developing innovation. This potential has a positive impact on economic growth, job creation and social progress. However, Artificial Intelligence systems may also pose risks to the respect of citizens' fundamental rights, such as those relating to discrimination and personal data protection, or even cause serious problems for the health or safety of citizens.

[13] It is the intention of the European Union that this Regulation "should meet a high level of protection of public interests, such as health and safety and the protection of funda-

which defines harmonised rules on this matter geared towards technological development, while at the same time offering a high level of protection of public interests and fundamental rights[14]. In particular, Artificial Intelligence systems classified as high-risk are subject to a set of requirements and specific obligations to ensure their proper functioning from a technical point of view and thus prevent damage to security and fundamental rights.

This means that high-risk systems must be solid, robust and accurate, i.e., efficient, of high quality, transparent, reliable and prepared to prevent and minimise behaviours that could cause damage, as well as wrong decisions or the generation of erroneous information. This determines the need to implement tests and evaluations that guarantee the accuracy of the technology used to ensure its robustness and solidity, as well as reliability, which is closely linked to ethical requirements.[15]

Article 15, which is the subject of this study, contains some of the mandatory technical requirements for high-risk systems that have to be fulfilled and must be subject to control. This control consists of a technical examination, prior to their placing on the market, demonstrating that these systems have been developed in accordance with the technical requirements laid down

mental rights, including democracy, the rule of law and the protection of the environment, as recognised and protected by Union law. In order to achieve this objective, rules should be laid down governing the placing on the market, putting into service and use of certain AI systems, thereby ensuring the proper functioning of the internal market and allowing such systems to benefit from the principle of free movement of goods and services. These rules should be clear and robust in protecting fundamental rights, support new innovative solutions, enable a European ecosystem of public and private actors to create AI systems in line with EU values and unlock the potential of digital transformation in all regions of the Union" (Recital 5).

[14] These standards are aligned with the Charter of Fundamental Rights of the European Union, the Union's international trade commitments, and should take into account the European Declaration on Digital Rights and Principles for the Digital Decade (2023/C 23/01) and the Ethical Guidelines for Trustworthy AI of the High Level Expert Group on Artificial Intelligence.

The seven basic principles that the European Commission considers necessary to establish and regulate for trustworthy AI are: human action and oversight; technical robustness and security; privacy and data management; transparency; diversity, non-discrimination and equity; social and environmental well-being; and accountability. These are set out in the *EU White Paper on Artificial Intelligence: A European approach to excellence and trust*, European Commission Communication COM (2020) 65 final, 19 February, p. 11; as well as in the Commission Communication COM (2019)168, 8 April, p. 4.

[15] We could think of different scenarios with negative effects on people, beyond the black boxes, less related to our study, we have the false positives and discriminatory decisions that result from biased systems or the unexpected or negative developments once the Artificial Intelligence system is put into production.

in harmonised standards set by the European standardisation bodies or in common specifications drawn up by the Commission or in other equivalent technical solutions generated by IT operators. Ultimately, high-risk systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle (Article 15).[16]

As can be seen, the three requirements have a direct relationship with each other, as robustness requires accuracy and cybersecurity, just as the latter requires the former. The solidity of the AI system aims to maintain the accuracy achieved at the beginning of the system's lifecycle when it is trained, tested and validated; and cybersecurity measures protect the system against possible attacks, thus ensuring its robustness and accuracy.

Similarly, the relationship with the quality of the data used for programming is also evident, as this is undoubtedly a major challenge when it comes to ensuring the proper functioning of AI systems. Access to quality data is fundamental to build robust and efficient intelligent systems, with important European initiatives, such as the EU Cybersecurity Strategy, the Digital Services Act and the Digital Markets Act, and the Data Governance Act, aiming to provide the appropriate infrastructure for building such systems.

On the other hand, it is necessary to refer to Royal Decree 817/2023 of 8 November, which establishes a sandbox for testing compliance with the proposal for a Regulation of the European Parliament and of the Council laying down harmonised standards on Artificial Intelligence. In its Article 11, located in Chapter 3 dedicated to "Test development, validation of compliance, monitoring and incidents", it establishes the requirements to be met during the development of tests that coincide with the ones we are dealing with here. Indeed, paragraph 1 states that: "Participation in the sandbox shall be aimed at complying, during the course of the test, with the implementation of the following requirements: h) Artificial Intelligence systems shall have been or be designed and developed so as to achieve, taking into account their intended purpose, an adequate level of accuracy, robustness and cybersecurity. These dimensions shall operate consistently throughout their lifecycle'.

In summary, it should be noted that in this paper we will address a number of issues, among which we highlight the following:

[16] In the words of Gamero Casado "These systems are not prohibited, but they are subject to a series of restrictions and to ex ante and ex post control mechanisms to guarantee the effective application of the Regulation. It is an orange light on the traffic light, since these systems can be implemented as long as the requirements established by the Regulation itself are met", *supra cit.* p. 279.

a) whether any changes or variations have occurred during the processing of the Regulation and, if so, their purpose and justification.
(b) how these technical quality requirements are specified to determine their compliance at appropriate quality levels.
c) how the Regulation ensures the quality of the data used for training, controls over training and model building, model evaluation metrics, maintenance and verification that any changes do not compromise the objective or original intent of the algorithm. The essential role of continuous monitoring of model performance metrics and detection of deviations from the concept.

Given the nature of this work, the methodology followed consisted of a comparative analysis of the different regulations applicable to AI, following the different versions of the proposals, as well as the doctrine that has been pronounced on the subject, with the aim of obtaining valid conclusions applicable to the international scientific community.

II. Development, processing and final content of Article 15

Article 15 establishes certain requirements for high-risk systems, requiring accuracy, robustness and cybersecurity throughout their lifecycle. Throughout its negotiation, this article has not been subject to substantial modifications except for the introduction of new paragraphs which, essentially, have integrated the development contained in the Recitals, which have been subject to a thorough revision in the final text. Throughout its processing with the General Guideline of 6 December 2022, the text remained practically unchanged, although the text of 14 June 2023 did introduce modifications. The final text has introduced changes in some precepts, even modifying the numbering.

In this regard, the former Recital 43 referred that "Requirements should apply to high-risk AI systems as regards the quality of data sets used, technical documentation and record-keeping, transparency and the provision of information to users, human oversight, and robustness, accuracy and cybersecurity. Those requirements are necessary to effectively mitigate the risks for health, safety and fundamental rights, as applicable in the light of the intended purpose of the system, and no other less trade restrictive measures are reasonably available, thus avoiding unjustified restrictions to trade". In the final text, Recital 46 states that "High-risk AI systems should only be placed on the Union market, put into service or used if they comply with certain mandatory requirements. Those requirements should ensure that high-risk

AI systems available in the Union or whose output is otherwise used in the Union do not pose unacceptable risks to important Union public interests as recognised and protected by Union law."

Also, in previous versions, recital 49 stated that high-risk AI systems should operate consistently throughout their lifecycle and present an adequate level of accuracy, robustness and cybersecurity in accordance with the generally recognised state of the art, with a duty to communicate to users the level of accuracy and the parameters used to measure it.

Technical robustness is defined as a key requirement for high-risk AI systems so that they must be resilient against harmful or undesirable behaviour that may result from the limitations of the systems or the environment in which they operate (e.g., errors, bugs, failures, inconsistencies, unexpected situations). In this regard, technical and organisational measures are required in both design and development to prevent or minimise harmful or undesirable behaviour. These measures include mechanisms to safely interrupt the operation of the system (fail-safe plans) in the presence of certain anomalies or when operation occurs outside certain predetermined limits.

On the other hand, in the last of the pre-final texts, recital 50, referring to the technical robustness of the system as a key requirement, talks about ensuring resilience to risks associated with the limitations of the system, as well as malicious actions that may compromise its security and lead to harmful or otherwise undesirable conduct. So the failure to protect against these risks could have security consequences or negatively affect fundamental rights, for example, due to wrong decisions being made or the AI system in question generating erroneous or biased output information.[17]

Amendment 86 proposed by the Parliament added that users should take measures to ensure that the possible trade-off between robustness and accuracy does not lead to discriminatory or negative outcomes for minority subgroups.[18]

These essential requirements of accuracy and robustness to effectively mitigate risks to health, safety and fundamental rights have a different but connected meaning. Accuracy is a quantitative measure of the relationship

[17] By biases we mean, following ISO (2006*): Statistics - Vocabulary and symbols - Part 1: General statistical terms and terms used in probability*, ISO, Tech. Rep. ISO 3534-1:2006. Available at: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standa rd/04/01/40145.html, "the degree to which a reference value deviates from the truth" or "a bias that favours or disadvantages a person, object or position" according to the *Ethical Guidelines…* cit. p. 48.

[18] In the June 2023 version it was recital 50, now in the final text this recital refers to cybersecurity.

between the intended purpose of the system and its performance from design to operation that provides insight into how the AI system performs in relation to its intended purpose and the data set it is working with. Therefore, when we talk about the accuracy of an AI model, we are talking about the proportion of correct predictions made by the model compared to the total number of predictions made. In other words, accuracy tells us how precise a model's predictions are, which is fundamental to the quality management system along with robustness, cybersecurity, transparency, data governance and monitoring.

The security and reliability of the AI system depends directly on its level of accuracy and robustness, which are inextricably linked because accuracy requires systems to be robust and therefore resistant to errors, failures and inconsistencies that may occur in the systems themselves or in the environment in which they operate, generally due to their interaction with natural persons or other systems (Article 15(3) of previous versions). Amendment 325 further clarified this paragraph 3(1) by stating that technical and organisational measures must be taken to ensure the resilience of high-risk AI systems by importing the terminology of the General Data Protection Regulation and its cross-cutting principle of proactive accountability.

This 3rd paragraph in relation to system robustness measures expressly referred to the adoption of technical redundancy solutions, such as backups or failure prevention plans, which amendment 326 expressly addressed to the relevant provider, with input from the user. These solutions give shape to the responsibility to take technical measures to ensure the robustness of the system in accordance with the intended purpose and taking into account foreseeable undesired results. This content in the final text has been moved to paragraph 4, which has also been reworded.

The Recitals develop the content of the articles and even go far beyond what is stated in the Regulation, hence their special interest as they are of great help in understanding the legislator's motivation. However, in the final text of the Regulation, numerous modifications have been made to the Recitals, to the point that all those that contained references to our subject of study have been completely changed in their content and enumeration.

Thus, Recital 59 with regard to Artificial Intelligence systems for law enforcement purposes makes an express reference to mandatory requirements when it states that 'In particular, if the AI system is not trained with high-quality data, does not meet adequate requirements in terms of its performance, its accuracy or robustness, or is not properly designed and tested before being put on the market or otherwise put into service, it may single out people in a discriminatory or otherwise incorrect or unjust manner'. Furthermore, Recital 61 states "In particular, to address the risks of potential biases, errors and opacity, it is appropriate to qualify as high-risk AI systems intended to be used by a judicial authority or on its behalf to assist judicial authorities in researching and interpreting facts and the law and in applying the law to a concrete set of facts".

As a novelty, recital 64 states that 'To mitigate the risks from high-risk AI systems placed on the market or put into service and to ensure a high level of trustworthiness, certain mandatory requirements should apply to high-risk AI systems, taking into account the intended purpose and the context of use of the AI system and according to the risk-management system to be established by the provider'. With regard to the adoption of measures, the rule is flexible, stating that providers "to comply with the mandatory requirements of this Regulation should take into account the generally acknowledged state of the art on AI, be proportionate and effective to meet the objectives of this Regulation". Furthermore, bearing in mind that a product is placed on the market or put into service only when it complies with the applicable EU harmonisation legislation, the provisions of the Regulation regarding the requirements to be met by high-risk systems refer to aspects other than those provided for in the EU harmonisation Acts and complement the sectoral regulation[19]. In this regard, the recital gives as an example "machinery or medical devices products incorporating an AI system might present risks not addressed by the essential health and safety requirements set out in the relevant Union harmonised legislation, as that sectoral law does not deal with risks specific to AI systems".

Further references to the mandatory requirements to be met by high-risk AI systems in order to effectively reduce the risks that may arise from their use where no other less trade-restrictive measures are available are found in Recital 66. Thus, "Requirements should apply to high-risk AI systems as regards risk management, the quality and relevance of data sets used, technical documentation and record-keeping, transparency and the provision of information to deployers, human oversight, and robustness, accuracy and cybersecurity".

As noted and explained below, data quality is crucial for the development of reliable and safe AI systems, and this quality must be maintained throughout the lifecycle of the system so that it does not degrade. In this respect,

[19] Commission Communication "Blue Guide" on the application of European product legislation of 2022", the general rule is that Union harmonisation legislation may be applicable to a product, as placing on the market or putting into service can only take place when the product complies with all applicable Union harmonisation legislation.

recital 67 in the final text, with an emphasis on techniques involving model training, relates data quality to ensuring that the system 'performs as intended and safely and it does not become a source of discrimination prohibited by Union law'.

Data requirements to be of sufficient quality, data management and governance practices to ensure that data sets for training, validation and testing are of high quality are regulated in the AIA with particular attention to the mitigation of biases that may negatively impact on fundamental rights or lead to discrimination prohibited by Union law, especially where output data influences input information for future transactions (feedback loops).

III. The requirements of an adequate level of accuracy and robustness

High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle (Article 15(1)). This adequate level shall be determined in light of their intended purpose and in accordance with the generally acknowledged state of the art (Recital 74). Earlier versions referred to maintaining an consistently level throughout their lifecycle "in accordance with the generally acknowledged state of the art."[20].

In the final text of the Regulation, a new paragraph 2 was introduced on the technical aspects of measuring the required levels of accuracy and robustness (paragraph 1) and for any other relevant performance metrics. It states that "s, the Commission shall, in cooperation with relevant stakeholders and organisations such as metrology and benchmarking authorities, encourage, as appropriate, the development of benchmarks and measurement methodologies"[21].

Recital 74, referred to above, explains how Union law on legal metrology, including Directives 2014/31/EU and 2014/32/EU of the European Parliament and of the Council, aims to ensure the accuracy of measurements and to contribute to transparency and fairness in commercial transactions. So in this context, in cooperation with relevant stakeholders and organisations, such as metrology and benchmarking authorities, the Commission should, as appropriate, encourage the development of benchmarks and measurement

[20] Recital 49 in the previous version which was intended to be amended by the Parliament with amendment 312 referring to the state of the art according to the specific market segment or scope.

[21] In previous versions 15.1.bis and 15.1.a.

methodologies for AI systems. In doing so, the Commission should take note of and collaborate with international partners working on relevant AI-related metrology and measurement indicators.[22]

The provider as the party responsible for the design, implementation, verification and validation of the AI system is primarily responsible for meeting these requirements throughout the entire lifecycle. It is therefore responsible for taking appropriate technical and organisational measures to ensure that the accuracy and robustness requirements of the system are met. Furthermore, within its scope of application, the user of the system assumes responsibilities that will materialise in specific technical and organisational measures.

In any case, the expected level of performance parameters should be stated in the instructions for use accompanying AI systems. Providers are encouraged to communicate such information to deployers in a clear and easily understandable manner, without misunderstanding or misleading statements (Recital 74)[23]. The requirement of the principle of transparency as a requirement for system quality requires that the instructions for use accompanying high-risk AI systems shall indicate the levels of accuracy of such systems, as well as the relevant parameters for assessing accuracy (Article 15(3)).[24]

[22] Similar terms were expressed in recital 49.

[23] The language should be clear and free from misunderstanding or misleading statements as proposed by the Parliament in its amendment 85 to recital 49 of the previous text. Recital 49 referred to the obligation to communicate to users the level of accuracy and the parameters used for measurement as a *sine qua non* for meeting the requirements in the design and development of the system. In this context, the expected level of performance metrics should be stated in the accompanying instructions for use, they should be described in the system documentation before designing further tests to be developed at the execution stage and this information should be communicated in a clear and easily understandable way, without misunderstandings or misleading statements.

[24] On the level of transparency of high-risk AI systems, as far as the contents of this paper are concerned, it should be noted that Article 13 refers to the characteristics, capabilities and limitations of the operation of the high-risk AI system, and in particular: (i) its intended purpose; (ii) the level of accuracy *(*including the parameters for assessing it*)*, robustness and cybersecurity referred to in Article 15 against which the high-risk AI system has been tested and validated and can be expected, as well as any known and foreseeable circumstances that may affect the expected level of accuracy, robustness and cybersecurity; (iii) any known or foreseeable circumstances, associated with the use of the high-risk AI system in accordance with its intended purpose or reasonably foreseeable misuse, which may give rise to risks to health and safety or fundamental rights as referred to in Article 9(2); (iv) where appropriate, the technical capabilities and characteristics of the high-risk AI system to provide relevant information to explain its output information; (v) where appropriate, its performance with regard to specific persons or groups of persons in relation to whom the system is intended to be used; (vi) where appropriate, input data specifications, or any other relevant information regarding the training,

The fulfilment of these requirements is designed with flexibility in the sense that technical solutions can be adopted from standards or other technical specifications or on the basis of general scientific or engineering knowledge at the discretion of the AI system provider concerned. Accordingly, AI system providers could choose how they want to meet the requirements, taking into account the state of the art and developments in that particular field[25]. Yes, it seems absolutely necessary that there is coordination of comparative evaluations to determine how the required standards should be measured.[26]

1. Metrics and system performance

The provider, as the responsible for the design, implementation, verification and validation of the AI system, must cover these requirements throughout the lifecycle of the system, as any aspect of the lifecycle can have an impact on the accuracy of the system. It is therefore the responsibility of the provider to take appropriate measures (both organisational and technical) to ensure that the minimum requirements set out in Article 15 are met.

However, while the accuracy of the system must be established or quantified throughout the system's lifecycle, certain steps are of particular relevance, namely the selection of data for training the system, which must be quality data. On the contrary, the use of erroneous, incomplete or biased data or false correlations have a negative impact on the system in terms of trust and reliability because they impede the goal of achieving "greater efficiency, accuracy, scale and speed of AI in making decisions and finding the best an-

validation and test data sets used, taking into account the intended purpose of the AI system; (vii) where appropriate, information to enable those responsible for deployment to interpret the output information from the high-risk AI system and to use it appropriately; (c) changes to the high-risk AI system and its operation predetermined by the provider at the time of the initial conformity assessment, if any; (d) the human surveillance measures referred to in Article 14, including technical measures put in place to facilitate the interpretation of output information from high-risk AI systems by those responsible for deployment; (e) the hardware and software resources required, the expected lifetime of the high-risk AI system and the maintenance and care measures required (including their frequency) to ensure the proper functioning of the high-risk AI system, including software updates.

[25] Recital 50 in its previous wording.

[26] Despite the existence of standardisation organisations to set standards, coordination is necessary, with the European AI Bureau convening national and international metrology and benchmarking authorities to provide non-binding guidance to address the technical aspects of measuring appropriate levels of accuracy and robustness.

swers"[27]. In the words of the European Parliament, given that "training data are often of questionable quality and are not neutral"[28], the "low quality" of data or procedures "could lead to biased algorithms, false correlations, errors, an underestimation of ethical, social and legal implications"[29] and ultimately result in decisions that negatively affect individuals and may even lead to (algorithmic) discrimination. This is an undesirable outcome, which the Regulation aims to avoid by regulating these technical requirements.

Article 10 of the Regulation regulates the data quality requirements. In this sense, data quality, as required by the standard, implies that the training, validation and test data sets are relevant, representative, free of errors and complete in terms of the intended purpose of the system. They must also have appropriate statistical properties, including with regard to the individuals or groups on which the high-risk AI system will initially be used. In particular, training, validation and test data sets should take into account, to the extent necessary for their intended purpose, the particular features, characteristics or elements of the specific geographical, behavioural or functional environment or context in which the AI system is intended to be used. In order to protect the rights of third parties against discrimination that could be caused by bias in AI systems, providers should also be able to process special categories of personal data, as a matter of essential public interest, to ensure that bias in high-risk AI systems is monitored, detected and corrected.

In order to improve data quality, which is essential for the proper functioning of the AI system, it would be interesting to follow the recommendation of authors such as Floridi[30], with a commitment to change the process of obtaining data. It would be a matter of abandoning Big Data in favour of data quality, and to this end it would be more relevant to stop working with huge amounts of data in order to choose smaller, but higher quality, sets. The higher quality is guaranteed through a careful selection of the data and with the reliability that this entails because the algorithms would be trained with better data that would no longer tend to be inaccurate, erroneous or contain biases.

[27] World Economic Forum 2018, p. 8.

[28] EUROPEAN PARLIAMENT (2017): Resolution of 14 March 2017 *on the fundamental rights implications of big data: privacy, data protection, non-discrimination, security and law enforcement* (2016/2225(INI)), point B. Available at: https://www.europarl.europa.eu/doceo/document/TA-8-2017-0076_ES.html

[29] European Parliament (2017), *cit.,* Recital m.

[30] Floridi, Luciano, "The Fight for Digital Sovereignty: What It Is, and Why It Matters, Especially for the EU" in *Philos. Technol*, no. 33, pp. 369-378 (2020). Available at: https://doi.org/10.1007/s13347-020-00423-6

This solution can be provided by the use of data generated by AI systems that meet the standards required by Article 10 of the Regulation.

In this regard, among the technical and organisational measures that system providers must establish, we find those aimed at selecting and evaluating precision metrics from the system design, as well as the system's quality controls, whose results depend on the verification and validation of these metrics[31]. In any case, the selection must be made based on several elements such as the purpose and the avoidance or mitigation of discrimination or bias.[32]

Performance metrics allow the evaluation of the performance of machine learning algorithms, quantifying the quality of the predictions, in order to mitigate the potential risks that the system represents. Different metrics provide a different perspective or view on the performance of the model, so it is important to choose the most appropriate metric for each task.

Thus, model accuracy (precision) is a commonly used metric that measures the proportion of correct predictions made by the model. This technique may be useful under certain assumptions and less so under others because significant class imbalance occurs and the accuracy does not provide a faithful representation of the model's performance. In any case, accuracy is a fundamental metric because it indicates the level of precision of the predictions made by a model compared to the total predictions. Thus, high accuracy guarantees reliable results and can make a difference in critical applications while low accuracy can have serious consequences. On the other hand, the calculation of accuracy may vary depending on the problem and the approach used.

As indicated above, in some cases it is convenient to use other metrics or even a combination of several, such as using precision with recall, specifically in cases of classification when it comes to assigning a label or category to an entry because precision measures the proportion of true positive predictions among all positive predictions and recall measures the proportion of true positive predictions among all true positive instances.

In the range of options is the F1 score metric that combines precision and recall, providing a single value that balances the trade-off between these two metrics, the score ranges from 0 to 1, with 0 being the worst performance

[31] US National Institute of Standards and Technology (NIST). AI Measurement and Evaluation. https://www.nist.gov/ai-measurement-and-evaluation; OECD.AI.CatalogueofTools&MetricsforTrustworthyAI. https://oecd.ai/en/catalogue/metrics, 2023; IEEE-StandardsAssociation. IEEEportfolioofAIStechnologyandimpactstandardsandstandardsprojects. https://standards.ieee.org/initiatives/autonomous-intelligence-systems/standards/

[32] Ensuring the accuracy of the model depends directly on the quality of the training data which must be representative of the intended purpose and free of bias, see Article 10 AIA.

and 1 being the best; the ROC Curve which plots the true positive rate (sensitivity) against the false positive rate (1-specificity) for different classification thresholds; or the ROC-AUC which provides a single value representing the overall performance of the model at all possible classification thresholds and the higher the value the better the performance.

The provider is also responsible for adopting other technical measures in terms of accuracy, such as the preparation of technical documentation containing all the information necessary for the correct implementation of the system and, where appropriate, the detection and communication of errors. The communication of metrics and system performance to the user[33] is, in any case, the responsibility of the provider, who must comply with the principle of transparency by providing all the information on the system as a quality indicator.

Once the metrics have been established, other organisational measures must be implemented to monitor the accuracy of the model and therefore whether it is working consistently, i.e., whether it is a solid and robust model. In this sense, metrics are an indicator of quality and a minimum to be met, so if it is not possible to guarantee them, human monitoring will be put in place.

Accuracy and robustness appear as inseparable requirements, as together with cybersecurity they are essential in high-risk Artificial Intelligence systems. Indeed, as indicated in previous pages, one of the objectives of system robustness is to ensure accuracy throughout the lifecycle of the system. Article 15 establishes the providers' responsibility to develop systems in such a way that they achieve an appropriate level of robustness, fit for purpose with the objective of mitigating the risks identified in the risk plan, which has to be maintained at an appropriate level and consistently throughout its lifecycle.

The aim is for systems to be resilient, resisting as much as possible to harmful or undesirable behaviour for various reasons such as limitations in the systems or the environment in which they operate, in particular due to their interaction with natural persons or other systems (Article 15(4), Recital 75).

The Regulation foresees that the provider puts in place measures such as technical redundancy solutions which may include back-up or fail-safe plans as tools to ensure the robustness and quality of the system. For example, data copying that ensures redundancy of models, algorithms, data, etc.; fail-safe mechanisms for components throughout the lifecycle; or the implementation

[33] It is the responsibility of the companies using the Artificial Intelligence system to know the level of accuracy and to have trained personnel in the organisation.

of an action plan when the system fails in order to recover the elements or reproduce the data (Article 15.4).

The failure to adopt protective measures against these risks could have security consequences or negatively affect fundamental rights, for example due to wrong decisions or wrong or biased output information provided by the AI system. Therefore, at the organisational level, the degradation of the system has to be monitored throughout the different stages of its lifecycle, controlling the constant quality of the data, preventing catastrophic oversights and taking into account feedback (Article 15.4 AIA).

These technical measures adopted to ensure the robustness of the system, through the prevention and minimisation of detrimental or undesirable system behaviours, must be documented to provide the user with the appropriate tools or mechanisms to observe, monitor and report different types of model degradation that exceed the documented reasonable limits for each robustness metric, in order to make them reproducible for correction. In addition, if the guaranteed robustness requirements change, it shall intervene to correct them in order to guarantee the metrics in the documentation.

In building robust AI systems, it is essential to analyse and understand the issues that arise in relation to the output data that come from the real world and the output data that are results provided by the model because of their direct impact on the level of accuracy of the system and, consequently, on its level of reliability.

This is a particularly problematic for high-risk AI systems that continue to learn after market introduction or put into service, as they change or modify their behaviour over time, creating scenarios where they introduce errors, failures or biased, unanticipated results that will influence the input data for future operations (feedback loops) and lead to a deterioration of their robustness and accuracy. In this regard, these systems shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures (Article 15.4).

System feedback is an iterative process in which the decisions and results of a model are continuously collected and used, with constant updating of training data, model parameters and system algorithms for the purpose of improving its performance. The risk of deterioration due to errors, biases and failures is high, more so if the model is trained not only on human-generated data but also on AI-generated data.

Therefore, systems that are retrained following an update process are forced to attend to feedback loops or feedback cycles that AI can differentiate between positive loops and negative loops, with the first type corresponding to those that generate accurate results that are aligned with users' expectations and preferences through positive comments left by people, which in turn reinforces the accuracy of future results. In another sense, AI negative feedback loops occur when AI models generate inaccurate results and users report these failures through a feedback loop that, in turn, attempts to improve the stability of the system by fixing the errors.

Also directly related to the robustness of the system is the desirability of establishing strategies to predict model failures that negatively affect fundamental rights or the safety of individuals in the use of AI systems. In short, using the words of the AIA, it is a matter of "preserving robustness as resistance to failures, errors or technical inconsistencies". In other words, ensuring the quality of the model from the beginning to the end of its lifecycle.

Possible degradation can occur while the model is being trained or when it is used for inference. Thus, problems can arise such as model deviation, which can be attributed to differences arising between what the model predicts and the truth. To avoid this type of degradation, measurement metrics such as variance, accuracy, precision, recall or bias are used. Ultimately, the tool to mitigate and control this type of differences or variations will be aimed at controlling this overlearning of the model.

On the other hand, another possible scenario occurs with model deviation over time, model drift, which occurs when the predictions of the learned model degrade due to changes in the environment. Therefore, predictive capabilities and efficiency decrease over time as the environment is changing and undergoes variations or alterations.

A third scenario occurs with data deviation over time, known as data drift or covariate shift, which happens when the input data of a model changes. This is the main reason why the precision of a model degrades over time (accuracy). Retraining the model can be a good solution to make it readapt to the changes and readjust to ensure its robustness.

As has been said, in those cases where the model training data are generated by AI, these problems that jeopardise the robustness of the system will increase due to the decrease in quality and their impact on the output results. So much so that this rapid development of generative AI has led to the study of the phenomenon known as model collapse, which is a degenerative process that negatively affects learned models because the generated data contaminates the training data set of the next model generation. In short, model collapse occurs because models are trained with AI-generated content instead of using human-generated content leading to a degradation of model quality. It would be a feedback loop because models trained with synthetic data will endlessly multiply errors, misinterpret data and outputs will be incorrect without taking into account less likely events because they fall outside the patterns.

The consequence would be large-scale data contamination.

Collapse can occur at different times. In the initial model when it starts to lose information about the tails of the training data distribution or, conversely, in the later model when it interweaves different modes of the original distributions and converges to a distribution that bears little or no resemblance to the original.

As for the reasons for the collapse of the model, we can establish two main categories. On the one hand, the statistical approximation error, which is the main error and is caused by the finite number of samples which, on the contrary, disappears as the sample count approaches infinity. On the other hand, functional approximation error occurs when certain models, such as neural networks, fail to capture the true underlying function to be learned from the data.

Ultimately, the feedback loops of AI models need to be robust and therefore of high quality.

As no one is unaware, accuracy is affected by the presence of biases, and it is essential that when metrics are chosen, a bias analysis is carried out to guarantee their reliability based on their impartiality. In this sense, the presence of biased, incomplete or noisy data, the excessive simplicity or complexity of the algorithm or the implicit or explicit biases carried by individuals must be understood in relation to the performance of the model in terms of systemic errors or deviations in results.

Bias admits different types because it can be the result of assumptions, preferences or limitations of the data, the algorithm, or the person involved in the modelling process. The analysis and approach to biases should not be done exclusively from a technical perspective and it is necessary to intervene in the human, social and institutional element where biases are embedded. Thus, three main categories of AI biases that need to be managed can be established:

- Systemic biases present in AI datasets, norms, practices, organisational processes throughout the AI lifecycle and, evidently, in the wider society that uses these systems.
- Computational and statistical biases in the datasets and algorithmic processes resulting from systemic errors due to non-representative samples.
- Human cognitive biases related to how a person or group perceives information from the AI system that will be used to make a decision or complete information being sought; as well as how we understand the purposes and functions of an AI system.

The model bias metrics can be applied at the data collection stage and at a later stage to assess the results after training the model, allowing for the detection of whether the predictions include biases.

2. Assessment of accuracy and robustness to ensure the quality of the system

The accuracy and robustness obtained has to be assessed through verification and subsequent validation. In other words, after confirming that the objectives have been met, they must be validated with objective evidence.

The objective is to determine how the system behaves and why it behaves the way it does and to be able to apply this information to improve its performance. The tools for assessing its performance are varied and are designed to measure results in order to determine whether or not the desired confidence threshold is reached. When measuring results, statistical methods are used to establish whether the desired confidence level or threshold is reached.

Validation involves testing the model using real data to verify the stability and effectiveness of the system. The model is tested on a separate dataset that has not been used during the training process to allow generalisation to new and unseen data. The cross-validation technique is very popular and involves splitting the dataset into multiple subsets that are used to train and test the model on different combinations of these subsets.

However, the quality of the model requires more than the evaluation of its performance through metrics and validation techniques, as its interpretability and fairness must be assessed. In this sense, we refer, on the one hand, to the ability to understand and explain the model's predictions, which is essential for building trust in AI systems, and, on the other hand, to the model's fairness, which implies that it does not discriminate against individuals or groups. Consequently, both the assessment of interpretability and the assessment of fairness are crucial for the validation of the model.

The evaluation of the model in terms of equity makes it possible to determine how the results of the model affect certain social groups defined according to attributes such as gender, race or age, among others, how the prediction values are distributed, and how the values of the performance metrics are among these groups. Based on the results obtained, it is evaluated how the differences, defined according to certain attributes, can be mitigated or corrected. Equity metrics allow for the assessment of performance levels to determine the reliability of the system as long as it does not risk becoming a source of discrimination through unfair results due to the presence of systemic biases that will disadvantage traditionally under-represented groups in the social reality. This result will be evidence of the good or bad performance of certain data sets that affect the robustness and reliability of the system.[34]

In conclusion, validation of AI systems is a complex process because it integrates the assessment of system performance in terms of accuracy and robustness, the identification and mitigation of biases in the data, and that privacy and security of personal data are respected. In addition, it also assumes that validated and documented technical and organisational measures have been implemented. However, it is by no means a closed process, but should be of a continuous nature, given the obligation to ensure the accuracy, robustness and cybersecurity of the system throughout the lifecycle of the AI system.

IV. Conclusions

High-risk AI systems place us in a scenario where decision-making can have a high impact on health, safety, and fundamental rights, so the results provided by the system must be precise or as accurate as possible, i.e., free of errors, inaccuracies and, of course, biases.

However, this priority objective is not at all easy in Artificial Intelligence models because the systems can be inaccurate or imprecise for different reasons, such as their continuous evolution over time and feedback processes, which directly and negatively affects them, generating a high degree of mistrust or unreliability.

System reliability is directly related to the principle or requirement of explainability, which, while it is a primary responsibility in AI and must be fulfilled by all those involved in the system's lifecycle, it is not easy to align with the requirement of technical accuracy. The system is explainable in terms of the level of explanation provided for how the system works and how the decision-making processes are carried out, but it also depends on how comprehensible these explanations are. Consequently, the higher the technical precision, the more complex the intelligent system, i.e., the less explicable.

On the other hand, reliable systems are robust systems, which operate accurately and consistently over time, with low levels of uncertainty and therefore with greater security and reliability. To ensure this technical solvency of the system, deployers should take appropriate technical and organisational

[34] Here, we refer to data quality analysis (Article 10) and data governance in terms of good data management and governance practices to ensure that training, validation and test datasets are of good quality.

measures to mitigate any risk arising from the breakdown of robustness and resilience. Artificial Intelligence expertise should be harnessed to design tools that adapt to different scenarios, i.e., specific measures that understand the complexity of each system and assess critical aspects such as performance, accuracy and algorithmic fairness.

Robustness and safety tests should be extensive to ensure that systems are reliable and safe in real environments and in the face of real hazards. This should be done without forgetting that an evaluation to detect biases must be conducted and thus ensure that AI systems are not biased and meet ethical requirements.

On the other hand, we consider that more detailed and specific requirements for impact assessment and technical tests that are mandatory for high-risk AI systems would be desirable in order to ensure the safety and quality of these systems.

In relation to the obligation of providers to demonstrate compliance with the established requirements, it seems appropriate to develop certification tools.

And, finally, international collaboration must go further in order to prevent and adequately respond to the technological challenges arising from the constant development of AI. Both in terms of global harmonised regulation and in terms of sharing best practices and learned lessons.

This analysis cannot be concluded without a reference to the guarantee of fundamental rights, in particular privacy, non-discrimination, and equal opportunities, which leads us to affirm that it is possible that in the not too distant future more specific and detailed requirements will have to be regulated to guarantee the protection of these rights.

CYBERSECURITY IN HIGH-RISK ARTIFICIAL INTELLIGENCE SYSTEMS IN ARTICLE 15 OF THE ARTIFICIAL INTELLIGENCE ACT

Marco Emilio Sánchez Acevedo

Lawyer. Lecturer and researcher at the Catholic University of Colombia.[1]

In this chapter, the reader will find an analysis of the "cybersecurity" obligation derived from article 15 of the new regulation with regard to Artificial Intelligence (hereinafter AI) systems classified as high risk. The methodology used corresponds to an analysis and interpretation of documents based on scanning, checking and interpretation. The purpose of this chapter is to determine which cybersecurity obligations must be met by Artificial Intelligence systems classified as high risk. To this end, three issues will be analysed; the first is related to resolving the question of whether cybersecurity applies to all Artificial Intelligence systems or only to high-risk ones, based on two questions of transcendental importance: on the one hand, the evolution and content of the cybersecurity obligation, and on the other, which systems are classified as high-risk and are required to comply with this obligation. It then addresses the European cybersecurity certification framework as an instrument to ensure compliance with the obligation set out in article 15 for high-risk Artificial Intelligence systems, and determines the cybersecurity obligations that must be fulfilled, including those implemented by public administrations and those related to critical infrastructures. Then, conformity assessments are addressed as an instrument for compliance with cybersecurity guarantees. Finally, conclusions are drawn.

[1] Marco Emilio Sánchez Acevedo is a lawyer in Colombia and Spain, teacher and researcher at the Catholic University of Colombia and collaborator in research projects at the University of Valencia - Spain, he holds a PhD in Technologies and Services of the Information Society, research line Public Law and Technologies; Master in Cybersecurity and National Cyberdefence, and Specialist in Administrative Law, Constitutional and e-Government. This research book chapter is the result of the work carried out by the author within the research project "Data analysis for spatial decision making and Artificial Intelligence for administration", linked to the research group Public Law and Technologies, category Colciencias A1, Catholic University of Colombia. Similarly, as a result of collaboration with the research group Régimen jurídico constitucional de las libertades, el gobierno abierto y el uso de las nuevas tecnologías (UVEG-GIUV2016-270), and the Project "Derechos y garantías públicas frente a las decisiones automatizadas y el sesgo y discriminación algorítmicas" 2023-2025 (PID2022-136439OB-I00) financed by MCIN/AEI/10.13039/501100011033/ ERDF, EU, as well as the MICINN national project, "Rights and guarantees against automated decisions in Artificial Intelligence, IoT, Big Data and robotics environments" (PID2019-108710RB-I00, 2020-2022) of the University of Valencia - Spain.

I. Does cybersecurity apply to all Artificial Intelligence systems? The Cybersecurity Obligation from the Proposal to the Final Approval

Cybersecurity has for several years now been one of the central elements in the development of information and communication technology projects in general terms. In 2013, the European Union's cybersecurity strategy (JOIN/2013) provided the Union's political response to the challenges related to cybersecurity. The first legal act in the field of Union cybersecurity was adopted in 2016, and corresponds to the Directive (EU) 2016/1148 of the European Parliament and of the Council. Directive (EU) 2016/1148 established a minimum legal framework to mitigate threats to networks and information systems, especially in the provision of essential services, as well as to seek tools to enable the continuity of services in the event of security incidents.

Information systems, networks and technologies in general are part of the central and daily life of citizens, companies and the government[2]. This generates a greater exposure to the set of threats that occur within the framework of the new cyberspace relations, and in the face of this, the challenges are amplified, the stakes must be understood from the search for answers to the risks, some of greater magnitude, others of lesser magnitude, but in any of the cases the need to face them arises. The use of Artificial Intelligence systems, as we have seen throughout the development of the new regulations analysed in this document, presents risks of a different nature. For this reason, the starting point for analysing the cybersecurity obligations imposed on Artificial Intelligence systems should be the Directive (EU) 2022/2555 of the European Parliament and of the Council, of 14 December 2022, on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) № 910/2014 and Directive (EU) 2018/1972 and repealing Directive (EU) 2016/1148 (NIS Directive 2).

Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity throughout the Union establishes a set of measures aimed at a high common level of cybersecurity throughout the Union and therefore incorporates a set of obligations regarding the need to adopt cybersecurity strategies, designate competent authorities, designate crisis management authorities, designate single points of contact and cybersecurity incident response teams; It also sets out measures to manage cybersecurity risks and reporting obligations, cy-

[2] On this topic, see Sánchez Acevedo, Marco Emilio et al., *El derecho y las tecnologías de la información y la comunicación (TIC)*, Universidad Católica de Colombia, Bogotá, 2015.

bersecurity information exchange obligations and oversight and enforcement obligations. The standard sets out a number of highly critical sectors, including the energy sector, which includes the electricity, heating, cooling, oil, gas, and hydrogen subsectors; the transport sector, which includes air transport by rail and water; the banking sector, financial market infrastructures, health, drinking water, digital infrastructure, waste water, digital infrastructure, ICT service management, business and public administration and space.

Effectively mitigating risks to health, safety and fundamental rights requires the imposition of a set of requirements that must be applied to Artificial Intelligence systems and that are linked to the quality of the data set used, technical documentation management, record keeping, information delivery and transparency, robustness, accuracy, human oversight and of course cybersecurity. Although the scope of the approved regulation covers high-risk AI systems, it is also true that the cybersecurity obligations are not exclusive to these systems, but to all information and communication technologies in a way that is proportionate to the purposes and interests pursued. The Regulation follows an approach based on the risks generated by the use of AI systems (i) an unacceptable risk, (ii) a high risk, and (iii) a low or minimal risk, for the case at hand in the present investigation, as mentioned above, the focus is only on high-risk AI systems. This does not imply that even if an AI information system does not fall into this category, it is still obliged to comply with cybersecurity standards.

Cybersecurity is the tool that will ensure that AI systems will cope with the different types of attacks that may occur and that will seek to exploit vulnerabilities. To ensure a level of cybersecurity appropriate to the risks, it is envisaged that providers of high-risk AI systems should take appropriate measures, taking also into account, as appropriate, the underlying ICT infrastructure. For the purpose of the content of this Chapter, Cybersecurity shall mean "the activities necessary to protect network and information systems, the users of such systems, and other persons affected by cyber threats" as referred to in Article 2(1) of Regulation (EU) 2019/881.[3]

[3] According to OECD work (see e.g. *Recommendation of the Council on Digital Security Risk Management for Economic and Social Prosperity in Digital Security Risk Management for Economic and Social Prosperity, OECD Recommendation and Companion Document*, OECD Publishing, Paris, 2015), 'cyber security' can be approached across dimensions 1) technology, when it focuses on the functioning of the digital environment (often called 'information security', 'computer security' or 'network security' by experts); 2) law enforcement or legal aspects (e.g. cybercrime); 3) national security, international stability, including aspects such as the role of ICTs with respect to intelligence, conflict prevention, warfare, cyber defence, etc., and 4) the economic and social dimension, covering wealth creation, innovation, growth, competitiveness and employment

1. Evolution, processing and final content of cybersecurity as a requirement in AI systems in the framework of the adopted proposal

Cybersecurity has been present throughout the activity of processing the draft regulation[4], both in the initial proposal, which refers to the need for AI systems to guarantee cybersecurity, and in the subsequent process, the agreement to incorporate cybersecurity obligations has been evident.

The proposal initially presented, in recital 43, starts by justifying that requirements, inter alia, for cybersecurity should apply to high-risk AI systems. In line with this, recital 49 incorporated that high-risk AI systems should, inter alia, have an adequate level of cybersecurity in accordance with the generally recognised state of the art. On the other hand, recital 51 justified it by stating that *'Cybersecurity plays a crucial role in ensuring that AI systems are resilient against attempts to alter their use, behaviour, performance or compromise their security properties by malicious third parties exploiting the system's vulnerabilities'*.

Already in the regulatory proposal, in particular Article 13 presented a cybersecurity obligation linked to transparency and communication of information to users, to the extent that high-risk systems shall be accompanied by instructions specifying the level of cybersecurity. In line with the above, Article 15 of the proposal presented that high-risk AI systems shall be designed and developed in such a way that, in view of their intended purpose, they achieve an adequate level of, inter alia, cybersecurity. In doing so, it linked cybersecurity to the intended purpose and by virtue of this, an adequate level of cybersecurity is envisaged. In the same vein, it stated that technical solutions aimed at ensuring the cybersecurity of high-risk AI systems should be appropriate to the relevant circumstances and risks.

One of the most relevant issues of the initial proposal is the presumption of compliance with certain cybersecurity-related requirements in high-risk AI systems that have been certified or for which a declaration of conformity has been issued under the cybersecurity scheme under Regulation (EU) 2019/881

in all economic sectors, individual freedoms, health, education, culture, democratic participation, science, entertainment and other dimensions of well-being where the digital environment drives progress.

[4] See the Artificial Intelligence Act (P9_TA(2023)0236), Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending Union legislation COM/2021/206, C9-0146/2021 and 2021/0106(COD) Retrieved from https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_ES.html

of the European Parliament and of the Council[5], insofar as the cybersecurity certificate or the declaration of conformity, or parts thereof, provide for these requirements.

In the amendments adopted by the European Parliament on 14 June 2023 on the proposal for a Regulation of the European Parliament and of the Council, which modify a couple of legislative acts of the Union (COM/2021/0206 - C9-0146/2021 - 2021/0106(COD) and some specific elements are incorporated, related to cybersecurity obligations and that strengthen the initial proposal, without making substantial changes, without prejudice to highlighting the following:

Amendment 17 incorporates a Recital 5a (new), which recognises, inter alia, "(…) cybersecurity concerns (…)".

Amendment 63, Recital (33a) incorporates, inter alia, a motivation to justify that '(…) Biometric and biometric-based systems provided for in Union law to enable cybersecurity and personal data protection measures should not be considered as posing a significant risk of harm to health, safety and fundamental rights'.

Amendment 64 adds to the initial proposal (34) "(…) In the case of critical infrastructure, it is appropriate to classify as high-risk the AI systems intended to be used as safety components in the management and operation of road traffic and the supply of water, gas, heating and electricity, since their failure or malfunctioning may put at risk the life and health of persons at large scale and lead to appreciable disruptions in the ordinary conduct of social and economic activities." Critical infrastructure security components, including critical digital infrastructures, are systems used to directly protect the physical integrity of critical infrastructure or the health and safety of people and property. A failure or malfunction of such components could directly lead to risks to the physical integrity of critical infrastructure and thus to risks to the health and safety of persons and property. Components intended to be used exclusively for cybersecurity purposes should not be considered as security components. Such security components include, for example, water pressure monitoring systems or fire alarm control systems in cloud computing centres.

Amendment 77 adds (43) that "Requirements should apply to high-risk AI systems as regards the quality of data sets used, technical documentation

[5] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (European Union Agency for Cybersecurity) and the certification of information and communication technology cybersecurity and repealing Regulation (EU) No 526/2013 (Cybersecurity Regulation) (OJ L 151, 7.6.2019, p. 1).

and record-keeping, transparency and the provision of information to users, human oversight, and robustness, accuracy and cybersecurity. Those requirements are necessary to effectively mitigate the risks for health, safety and fundamental rights, as applicable in the light of the intended purpose of the system, and no other less trade restrictive measures are reasonably available, thus avoiding unjustified restrictions to trade".

Amendment 85 adds (49) "High-risk AI systems should perform consistently throughout their lifecycle and meet an appropriate level of accuracy, robustness and cybersecurity in accordance with the generally acknowledged state of the art (…)".

In amendment 87 it is added (51) "Cybersecurity plays a crucial role in ensuring that AI systems are resilient against attempts to alter their use, behaviour, performance or compromise their security properties by malicious third parties exploiting the system's vulnerabilities. Cyberattacks against AI systems can leverage AI specific assets, such as training data sets (e.g. data poisoning) or trained models (e.g. adversarial attacks), or exploit vulnerabilities in the AI system's digital assets or the underlying ICT infrastructure. To ensure a level of cybersecurity appropriate to the risks, suitable measures should therefore be taken by the providers of high-risk AI systems (…)".

Amendment 101 (60g) is added "In view of the nature and complexity of the value chain for AI systems, it is essential to clarify the role of actors contributing to the development of such systems. (…) in particular, foundational models should assess and mitigate potential risks and harms through appropriate design, testing and analysis, implement data governance measures - in particular a bias assessment - and meet technical design requirements that ensure appropriate levels of performance, predictability, interpretability, co-readability, security and cyber-security, as well as comply with environmental standards. (…)'.

Amendment 110 adds (65) "In order to carry out third-party conformity assessment for AI systems intended to be used for the remote biometric identification of persons, notified bodies should be designated under this Regulation by the national competent authorities (…) with regard to their independence, competence and absence of conflict of interest, as well as minimum cyber-security requirements. (…)".

Amendment 115 adds (69) "(…) the Commission should take into account cybersecurity risks and risks linked to hazards. In order to maximise the availability and use of the database by the public, the database and the information made available through it should comply with the requirements set out in Directive 2019/882".

(ii) The level of accuracy, robustness and cybersecurity referred to in Article 15 against which the high-risk AI system has been tested and validated and can be expected to perform, as well as the clearly known or foreseeable circumstances that could affect the expected level of accuracy, robustness and cybersecurity (…)".

Amendment 321 adds "(…) 1. High-risk AI systems shall be designed and developed following the principle of safety by design and by default. In view of their intended purpose, they shall achieve an adequate level of accuracy, robustness, security and cybersecurity and perform consistently in these respects throughout their lifecycle (…)".

Amendment 323 adds to Article 15 - paragraph 1b (new) 1b: "In order to address any emerging issues in the internal market in relation to cybersecurity, the European Union Agency for Cybersecurity (ENISA) shall collaborate with the European Committee on Artificial Intelligence as set out in Article 56(2)(b)".

In amendment 399 it is proposed to add to Article 28b (new) as regards Obligations of the provider of a foundation model "(…) 1. Before placing it on the market or putting it into service (…) 2. For the purposes of paragraph 1, the provider of a foundation model shall: (a) demonstrate, through appropriate design, testing and analysis, the detection (…) (c) design and develop the foundation model in order to achieve throughout its lifecycle adequate levels of performance, predictability, interpretability, correctness, security and cybersecurity assessed by appropriate methods, such as model evaluation with the involvement of independent experts, documented analysis and comprehensive testing during conceptualisation, design and development; (…)".

Amendment 401 seeks to add to Article 29 - paragraph 1a (new) 1a: "To the extent that implementers exercise control over the high-risk AI system, they shall: I(…); iii) ensure that relevant and appropriate robustness and cybersecurity measures are regularly monitored for effectiveness and periodically adjusted or updated".

Amendment 423 seeks to add to Article 33 - paragraph 2 "Notified bodies shall meet the organisational requirements, as well as quality, resource and process management requirements, necessary for the performance of their functions, as well as the minimum cybersecurity requirements established for public administration entities identified as operators of essential services" in accordance with Directive (EU) 2022/2555.

Amendment 505 proposes an Article 53a (new) "Modalities and operation of controlled test sites for AI 1(…) 2. The Commission shall be empowered to adopt delegated acts in accordance with the procedure referred to in Article 73 not later than twelve months after the entry into force of this Regulation and shall ensure that: (…) (h) sandboxes shall facilitate the development of tools and infrastructures for testing, benchmarking, assessing and explaining the dimensions of AI systems relevant to sandboxes, such as accuracy, robustness and cybersecurity, as well as minimising risks to fundamental rights, the environment and society as a whole. 3. (…);".

Amendment 532 proposes a new Article 57a "Composition of the Management Board 1. The Management Board shall be composed of the following members: (…) (d) a representative of the European Union Agency for Cyber Security (ENISA); (…) Each representative of a national supervisory authority shall have one vote. The representatives of the Commission, the EDPS, ENISA and the FRA shall not have the right to vote. Each member shall have one alternate. The appointment of the members and alternates of the management board shall take into account the need for gender balance. The members of the management board and their alternates shall be made public (…)'.

Amendment 557 adds to the initial proposal "(…) 4. Member States shall ensure that the supervisory authority has adequate technical, financial and human resources and infrastructure for the effective performance of its tasks under this Regulation. In particular, the national supervisory authority shall have at its permanent disposal sufficient staff whose skills and expertise shall include a thorough knowledge of Artificial Intelligence, data and data computing technologies, personal data protection, cybersecurity, competition law, risks to fundamental rights, health and safety, and knowledge of existing legal rules and requirements. (…)".

In amendment 559 corresponding to Article 59 - paragraph 4b (new) 4b "National supervisory authorities shall satisfy the minimum cybersecurity requirements for public administration entities identified as operators of essential services under Directive (EU) 2022/2555".

In amendment 640 a proposal is made to add Article 70 - paragraph 1a (new) 1a. "The authorities involved in the application of this Regulation in accordance with paragraph 1 shall minimise the amount of data requested for disclosure to the data strictly necessary for the perception of risk and the assessment of that risk. They shall delete the data as soon as they are no longer necessary for the purpose for which they were requested. They shall put in place appropriate and effective cybersecurity, technical and organisational measures to protect the security and confidentiality of the information and data obtained in the performance of their tasks and activities.

Amendment 755 corresponding to Annex IV - paragraph 1 - point 2 - point g adds "the validation and test procedures used, including information about the validation and test data used and their main characteristics; the parameters used to measure accuracy, robustness and compliance with other relevant requirements laid down in Chapter 2 of Title III, as well as potentially discriminatory effects; test logs and all test reports dated and signed by the responsible persons, in particular with regard to the predetermined changes referred to in point (f)".

Thus, it is clear that cybersecurity, at this stage of the process, occupied several of the debates and becomes one of the essential elements to be regulated in the regulatory proposal.

Finally, the already approved regulation establishes a set of obligations in a comprehensive manner in the area of cybersecurity, among other issues, it highlights; i) not considering biometric systems intended to be used solely for the purpose of enabling cybersecurity and personal data protection measures to be high-risk systems (Recital 54); (ii) AI systems intended to be used as security components in the management and operation of critical digital infrastructures listed in Annex I(8) to Directive (EU) 2022/2557 should be classified as high risk, however, components intended to be used solely for cybersecurity purposes should not be considered as security components (Recital 55); (iii) requirements should apply to high-risk AI systems as regards, inter alia, risk management and cybersecurity (Recital 66); high-risk AI systems should operate consistently throughout their lifecycle and achieve an adequate level of, inter alia, cybersecurity, in the light of their intended purpose and in accordance with the generally recognised state of the art (74); Cybersecurity is essential to ensure that AI systems are resilient to actions by malicious third parties who, exploiting vulnerabilities in the system, seek to alter their use, behaviour or operation or to compromise their security properties, and to ensure a level of cybersecurity appropriate to the risks, providers of high-risk AI systems should take appropriate measures, such as security controls, also taking into account, where appropriate, the underlying ICT infrastructure (recital 76); high-risk AI systems falling within the scope of Regulation 2022/0272, in accordance with Article 8 of Regulation 2022/0272[6], may demonstrate compliance with the cybersecurity requirement by meeting the essential cybersecurity requirements set out in Article 10 and in Annex I of Regulation 2022/0272 and should be considered as compliant with the cybersecurity requirements set out in the EU declaration of conformity or parts thereof issued in accordance with Regulation 2022/0272 (recital

[6] European Economic and Social Committee, Opinion of the European Economic and Social Committee on the proposal for a Regulation of the European Parliament and of the Council on horizontal requirements for cybersecurity for products with digital elements and amending Regulation (EU) 2019/1020 (Rapporteur, Mensi, Maurizio), Retrieved from https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52022AE4103

77); providers of general purpose AI models with systemic risks should assess and mitigate potential systemic risks (114 and 115); high-risk AI systems that have been certified or for which a declaration of conformity has been issued under a cybersecurity regime contained in Regulation (EU) 2019/881 of the European Parliament and of the Council, comply with the cybersecurity requirement of this Regulation to the extent that the cybersecurity certificate or declaration of conformity or parts thereof cover the cybersecurity requirement.

In the area of cybersecurity, the approved regulation incorporates a set of direct obligations in the articles, all linked to high-risk systems, which must be addressed in parallel to the obligations of transparency and provision of information to deployers (Article 13) and accuracy, robustness and cybersecurity (Article 15).

The first obligation is linked to the fact that *'(…) 2. high-risk AI systems shall be accompanied by instructions for use in an appropriate digital format or otherwise that include concise, complete, correct and clear information that is relevant, accessible and comprehensible to deployers (…) the instructions for use shall contain at least (…) (i) its intended purpose; (ii) the level of accuracy, including its metrics, robustness and cybersecurity referred to in Article 15, against which the high-risk AI system has been tested and validated and which can be expected, and any known and foreseeable circumstances that may have an impact on that expected level of accuracy, robustness and cybersecurity'.* Similarly, *'(…) High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle. (…) High-risk AI systems shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities. The technical solutions aiming to ensure the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks. The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws'.*

As regards the requirements for notified bodies, it establishes the duty of these bodies to comply with suitable cybersecurity requirements (Article 31).

Similarly, Article 42 "*Presumption of conformity with certain requirements" in paragraph 2 (…) "High-risk AI systems that have been certified or for which a statement of conformity has been issued under a cybersecurity scheme pursuant to Regulation (EU) 2019/881 and the references of which have been published in the Official Journal of the European Union shall be presumed to comply with the cybersecurity requirements set out* *in Article 15 of this Regulation in so far as the cybersecurity certificate or statement of conformity or parts thereof cover those requirements*".

With regard to the obligations of providers of general purpose AI models with systemic risk, providers of these models shall ensure that an adequate level of cybersecurity protection is in place for the general purpose AI model with systemic risk and the physical infrastructure of the model.

As regards implementing acts to avoid fragmentation in the Union, an obligation is laid down for the Commission to adopt implementing acts specifying the detailed arrangements for the establishment, development, implementation, operation and monitoring of AI regulatory sandboxes. In line with this, the implementing acts referred to in Article 58(1) shall ensure that the AI regulatory sandboxes facilitate the development of tools and infrastructures for testing, benchmarking, evaluating and explaining the dimensions of AI systems relevant for regulatory learning, such as accuracy, robustness and cybersecurity, as well as measures to reduce risks to fundamental rights and society as a whole.

With the creation of the European Artificial Intelligence Board (regulated in Article 65), it is given obligations to cooperate, inter alia, in the field of cybersecurity (Article 66(h)). In the same vein, an Advisory forum (Article 67) is established to provide expertise and advice to the Board and the Commission. The Fundamental Rights Agency, the European Union Agency for Cybersecurity (ENISA), the European Committee for Standardisation (CEN), the European Committee for Electrotechnical Standardization (CENELEC) and the European Telecommunications Standards Institute (ETSI) will be permanent members of the advisory forum (Article 67(5)).

Finally, as regards the technical documentation incorporated in Annex IV, it states that "The technical documentation referred to in Article 11(1) shall include at least the following information, applicable to the relevant AI system: (…) (h) (h) the cybersecurity measures taken".

Thus, the rule contains a set of direct obligations and others that refer us to the application of obligations contained in other rules, but always linked to high-risk AI systems.

2. Cybersecurity in high-risk AI systems

Cybersecurity obligations are linked to high-risk AI systems contained in Annex II and III of the new Regulation[7]. In the same sense, the product of which the AI system is a safety component according to point (a), or

[7] On this issue, see the section on high-risk systems developed above.

the AI system itself as a product, shall be considered as high risk if it must undergo a conformity assessment by an independent body for its placing on the market or putting into service in accordance with the Union harmonisation legislation listed in Annex I, and therefore with cybersafety obligations as referred to in Article 15: safety of toys[8], recreational craft and personal watercraft[9], lifts and safety components for lifts[10], equipment and protective systems intended for use in potentially explosive atmospheres[11], placing on the market of radio equipment[12], marketing of pressure equipment[13], cableway installations[14], personal protective equipment[15], gas appliances[16], medical devices[17], in vitro diagnostic medical devices[18], civil aviation safety[19], type-ap-

[8] Directive 2009/48/EC of the European Parliament and of the Council on the safety of toys, 18 June 2009, p. 1 (OJ L 170, 30.6.2009).

[9] Directive 2013/53/EU of the European Parliament and of the Council on recreational craft and personal watercraft and repealing Directive 94/25/EC, 20 November 2013, p. 90 (OJ L 354, 28.12.2013).

[10] Directive 2014/33/EU of the European Parliament and of the Council on the harmonisation of the laws of the Member States relating to lifts and safety components for lifts, 26 February 2014, p. 251 (OJ L 96, 29.3.2014).

[11] Directive 2014/34/EU of the European Parliament and of the Council on the harmonisation of the laws of the Member States concerning equipment and protective systems intended for use in potentially explosive atmospheres, 26 February 2014, p. 309 (OJ L 96, 29.3.2014).

[12] Directive 2014/53/EU of the European Parliament and of the Council on the harmonisation of the laws of the Member States relating to the placing on the market of radio equipment and repealing Directive 1999/5/EC, 16 April 2014, p. 62 (OJ L 153, 22.5.2014).

[13] Directive 2014/68/EU of the European Parliament and of the Council on the harmonisation of the laws of the Member States concerning the placing on the market of pressure equipment, 15 May 2014, p. 164 (OJ L 189, 27.6.2014).

[14] Regulation (EU) 2016/424 of the European Parliament and of the Council on cableway installations and repealing Directive 2000/9/EC, 9 March 2016, p. 1 (OJ L 81, 31.3.2016).

[15] Regulation (EU) 2016/425 of the European Parliament and of the Council on personal protective equipment and repealing Council Directive 89/686/EEC, 9 March 2016, p. 51 (OJ L 81, 31.3.2016).

[16] Regulation (EU) 2016/426 of the European Parliament and of the Council of 9 March 2016 on appliances burning gaseous fuels and repealing Directive 2009/142/EC (OJ L 81, 31.3.2016, p. 99).

[17] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC (OJ L 117, 5.5.2017, p. 1).

[18] Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU (OJ L 117, 5.5.2017, p. 176).

[19] Regulation (EC) No 300/2008 of the European Parliament and of the Council of 11

proval and market surveillance of two and three-wheel vehicles and quadricycles[20], type-approval and market surveillance of agricultural and forestry vehicles[21], marine equipment[22], interoperability of the rail system in the European Union[23], type-approval and market surveillance of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles[24], type-approval of motor vehicles and their trailers[25], common rules in the field of civil aviation.[26]

March 2008 on common rules in the field of civil aviation security and repealing Regulation (EC) No 2320/2002 (OJ L 97, 9.4.2008, p. 72).

[20] Regulation (EU) No 168/2013 of the European Parliament and of the Council of 15 January 2013 on type-approval and market surveillance of two- or three-wheel vehicles and quadricycles (OJ L 60, 2.3.2013, p. 52).

[21] Regulation (EU) No 167/2013 of the European Parliament and of the Council of 5 February 2013 on type-approval and market surveillance of agricultural and forestry vehicles (OJ L 60, 2.3.2013, p. 1).

[22] Directive 2014/90/EU of the European Parliament and of the Council of 23 July 2014 on marine equipment and repealing Council Directive 96/98/EC (OJ L 257, 28.8.2014, p. 146).

[23] Directive (EU) 2016/797 of the European Parliament and of the Council of 11 May 2016 on the interoperability of the rail system in the European Union (OJ L 138, 26.5.2016, p. 44).

[24] Regulation (EU) 2018/858 of the European Parliament and of the Council of 30 May 2018 on type-approval and market surveillance of motor vehicles and their trailers, and of systems, components and separate technical units intended for such vehicles, amending Regulations (EC) No 715/2007 and (EC) No 595/2009 and repealing Directive 2007/46/EC (OJ L 151, 14.6.2018, p. 1).

[25] Regulation (EU) 2019/2144 of the European Parliament and of the Council of 27 November 2019 concerning type-approval requirements for motor vehicles and their trailers, together with their systems, components and separate technical units intended for such vehicles, with regard to their general safety and the protection of their occupants and vulnerable road users, amending Regulation (EU) 2018/858 of the European Parliament and of the Council and repealing Regulations (EC) No 78/2009, (EC) No 79/2009 and (EC) No 661/2009 of the European Parliament and of the Council and Regulations (EC) No 631/2009, (EC) No 406/2009, (EC) No 406/2009 and (EC) No 406/2009 of the European Parliament and of the Council and Regulations (EC) No 631/2009, (EC) No 406/2009 and (EC) No 406/2009 of the European Parliament and of the Council.º 78/2009, (CE) N.º 79/2009 y (CE) N.º 661/2009 del Parlamento Europeo y del Consejo y los Reglamentos (CE) N.º 631/2009, (UE) N.º 406/2010, (UE) N.º 672/2010, (UE) N.º 1003/2010, (UE) N.º 1005/2010 de la Comisión, (UE) N.No 1008/2010, (EU) No 1009/2010, (EU) No 19/2011, (EU) No 109/2011, (EU) No 458/2011, (EU) No 65/2012, (EU) No 130/2012, (EU) No 347/2012, (EU) No 351/2012, (EU) No 1230/2012 and (EU) 2015/166 (OJ L 325, 16.12.2019, p. 1).

[26] Regulation (EU) 2018/1139 of the European Parliament and of the Council of 4 July 2018 on common rules in the field of civil aviation and establishing a European Union Aviation Safety Agency, amending Regulations (EC) No 2111/2005, (EC) No 1008/2008, (EU) No 996/2010, (EU) No 376/2014 and Directives 2014/30/EU and 2014/53/EU of the Europe-

Although cybersecurity obligations are incorporated, it is also true that the regulation directs cybersecurity to the application of existing standards and to the level appropriate to the intended purpose and circumstances. Accordingly, cybersecurity standards provided by the European Union should be applied, among others, and in particular Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity throughout the Union[27], amending Regulation (EU) № 910/2014 and Directive (EU) 2018/1972. and repealing Directive (EU) 2016/1148 (NIS Directive 2), and Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (European Union Agency for Cybersecurity) and Information and Communication Technology Cybersecurity Certification and repealing Regulation (EU) № 526/2013 ("Cybersecurity Regulation"), Regulation 2022/0272[28], among others.

The latter contains the requirements of the European Cybersecurity Certification Scheme, understood as the "complete set of provisions, technical requirements, standards and procedures established at Union level that apply to the certification or conformity assessment of specific ICT products, services and processes", which is articulated at national level through the "National Cybersecurity Certification Scheme" understood as the "complete set of provisions, technical requirements, standards and procedures developed and adopted by a national public authority, and which apply to the certification or conformity assessment of ICT products, services and processes falling within the scope of that specific scheme", and is materialised in practical terms as the "European Cybersecurity Certificate", which corresponds to the "document issued by the relevant body certifying that a given ICT product,

an Parliament and of the Council, and repealing Regulations (EC) No 552/2004 and (EC) No 216/2008 of the European Parliament and of the Council and Council Regulation (EEC) No 3922/91 (OJ L 212, 4.8.2018, p. 1).No 552/2004 and (EC) No 216/2008 of the European Parliament and of the Council and Council Regulation (EEC) No 3922/91 (OJ L 212, 22.8.2018, p. 1), as regards the design, production and placing on the market of aircraft referred to in Article 2(1)(a) and (b) thereof, when they are unmanned aircraft and their engines, propellers, parts and appliances to control them remotely.

[27] Rule implemented in articulation with Regulation (EU) 2016/679 of the European Parliament and of the Council) and in Directive 2002/58/EC of the European Parliament and of the Council.

[28] Opinion of the European Economic and Social Committee on the proposal for a Regulation of the European Parliament and of the Council on horizontal requirements for cybersecurity for products with digital elements and amending Regulation (EU) 2019/1020. Retrieved from https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:52022AE4103

service or process has been assessed to verify that it meets the specific security requirements set out in a European Cybersecurity Certification Scheme".

II. The European Cybersecurity Certification Framework as an assurance tool for high-risk AI systems

The European Framework for the certification of cybersecurity establishes a scheme for certification and confirmation that products, processes and services associated with information and communication technologies have been assessed and comply with requirements to protect the authenticity, integrity, availability and confidentiality of data whether stored, transmitted or processed, or any service or function that can be accessed during the lifecycle of products, services, and processes.

Based on article 47 of Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019, it could be proposed that the EU's evolving work programmes should include information and communication technology products, services or processes, and in particular AI systems classified as high-risk, and thus, that AI systems should have independent security certification. This is justified by applicable EU law or policies, particularly new ones on Artificial Intelligence systems, market demand and the evolution of cyber threats in Artificial Intelligence environments.

The design of European Cybersecurity Certification schemes addresses several key objectives to ensure security in the lifecycle of Information and Communication Technology (ICT) products, services or processes. These objectives include protection against unauthorised access, preservation of data integrity and availability, proper management of authorised access, detection and documentation of known vulnerabilities, logging and verification of access and usage activities, elimination of vulnerabilities in products and rapid restoration of services in case of incidents. In addition, the importance of security by default and by design is emphasised, as well as the delivery of products and services with up-to-date and secure software and hardware, with mechanisms for security updates.

European cybersecurity certificates may specify one or more of the following assurance levels for ICT products, services and processes: 'basic', 'substantial' or 'high'. The assigned assurance level should reflect the risk associated with the intended use of an ICT product, service or process, considering both the likelihood and potential impact of a cybersecurity incident. A European cybersecurity certificate or an EU declaration of conformity, designated as "basic" level, ensures that ICT products, services and processes comply with security requirements, minimising the known risks of cyber incidents and cyber attacks. The assessment includes at least a review of technical documentation or equivalent assessment activities. In the case of a "substantial" level certificate, it ensures that ICT products, services and processes, comply with security requirements, minimising known cybersecurity risks and attacks by resource-constrained actors. The assessment involves review to demonstrate the absence of known vulnerabilities and verification of the correct implementation of security functionalities. Finally, a "high" level certificate provides assurance that ICT products, services and processes comply with security requirements, minimising the risk of sophisticated cyber-attacks by actors with considerable capabilities and resources.

Given the above, the question arises as to what level of assurance will be required for AI systems? The answer could lie in the purpose of the AI system, the rule provides that "High-risk AI systems shall be designed and developed in such a way as to achieve, in light of their intended purpose, an adequate level of (…) cybersecurity". In practice, any high-risk AI systems, given the levels of risk they pose to rights, should be integrated into the "high" cybersecurity certificate model. Downgrading the category to lower levels would imply a substantial contradiction with the standard itself as it would be the minimisation of risks of cyber-attacks, cyber-incidents and a minor revision in terms of assessment technique or activities.

Similarly, high-risk AI systems should comply with the obligations arising from Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022, as regards the adoption of national cybersecurity strategies, incorporating the specific section on high-risk AI systems and designating or establishing competent authorities, cybersecurity crisis management authorities, single points of contact on cybersecurity and cybersecurity incident response teams. Similarly, the integration of these into the models designed for cybersecurity risk management and notification obligations, and the identification of entities whose type falls under Annex I or II; as well as for entities identified as critical under Directive (EU) 2022/2557. Moreover, obligations regarding the exchange of information on cybersecurity and those arising from oversight, supervision and control activities.

1. Cybersecurity certifications in high-risk AI systems, mandatory or voluntary?

From the cybersecurity obligation imposed in Article 15 of the new AI Act, it follows that the cybersecurity rules previously adopted in Union law apply, in particular Article 56 of Regulation (EU) 2019/881 [29] of the European Parliament and of the Council, of 17 April 2019, as regards cybersecurity certifications and in particular the second paragraph of that provision which states that cybersecurity certification shall be voluntary, unless otherwise provided for in Union or Member State law.

For high-risk AI systems in which the cybersecurity obligation set out in Article 15 has been incorporated, it should then, on the one hand, have a specific European cybersecurity certification scheme for AI systems that meets the elements set out in Article 54 and, on the other hand, become mandatory under EU law. If so, they should have such certification assessed every two years, and, at the same time the Commission should determine on the basis of the results of that assessment the products, services and processes covered by the mandatory AI cybersecurity certification scheme. Similarly, manufacturers or providers of certified or self-assessed ICT products, services and processes must provide the complementary cybersecurity information referred to in Article 55, including guidance on secure product maintenance, support period, updates, vulnerability information.

The development of cybersecurity certificates in AI systems should also take into account the peer review contained in Article 59, which aims to "achieve equivalent standards across the Union for European cybersecurity certificates issued and EU declarations of conformity".

The cybersecurity certification scheme for high-risk AI systems must be brought into line with the internal rules of each of the States, and it is certain that, in some more than others, the adaptations will be considerable. As an example, and only by way of illustration, see the need to adapt the Spanish rules contained in the Order PRE/2740/2007, of 19 September, which approves the Regulation on the Evaluation and Certification of Information Technology Security, whose purpose is "the articulation of the Certification Body (CB) of the National Information Technology Security Evaluation and Certification Scheme (NITES) within the scope of action of the National Cryptologic Centre, according to the provisions of Law 11/2002, of 6 May, regulating the National Intelligence Centre, and Royal Decree 421/2004, of 12 March, which regulates the National Cryptologic Centre, respectively". As well as the recent Royal Decree 311/2022, of 3 May, which regulates the National Security Scheme and whose purpose is to regulate the National Security

[29] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (European Union Agency for Cybersecurity) and the certification of information and communication technology cybersecurity and repealing Regulation (EU) No 526/2013 ("Cybersecurity Regulation").

Scheme, established in article 156.2 of Law 40/2015, of 1 October, on the Legal Regime of the Public Sector.

Information security seeks to ensure that an organisation can achieve its objectives and perform its functions through the use of information systems. To achieve this, fundamental principles must be followed, including viewing security as a holistic process, risk-based management, addressing prevention, detection, response and preservation, establishing lines of defence, maintaining continuous vigilance, conducting periodic reassessments and differentiating responsibilities. These principles are key to establishing an effective and robust approach to information security.

2. Cybersecurity obligations in high-risk AI systems

Cybersecurity involves the development of capabilities around the process of identification, protection, detection, response and recovery. Minimising vulnerabilities and ensuring that risks do not materialise requires a process of planning and understanding of cybersecurity from the beginning and throughout the development process of high-risk AI systems. Cybersecurity should be conceived as a comprehensive set of actions at each stage of preparation, development, deployment and control of the AI system. Both the European cybersecurity framework and the national cybersecurity framework of each EU Member State, in accordance with technical standards, include, integrate, and develop the best-known cybersecurity models. Andrew S. Tanenbaum pointed out that "the good thing about standards is that there are so many to choose from"; the ISM 3 information security management systems maturity model, the ISO 27001 information security management system, the White Community Cybersecurity Maturity Model (CCSMM), the NIST cybersecurity Framework, among others, which develop in one way or another the different stages of identifying, protecting, detecting, responding to and recovering from cyber attacks.

The cybersecurity obligation contained in Article 15 implies the strengthening of the cybersecurity function, based on the application of international technical standards, for those who are included as obliged subjects to comply with the obligations derived from the AIA and especially for high-risk AI systems. Security should be an initial principle of such projects and consequently activities should be established to manage assets, identify the service environment, have identified levels of cybersecurity governance, have programmes for risk assessment, identify the strategy to manage risks, identify the risk management model in the supply chain; the implementation of protection activities e.g., identity management, access controls, information protection procedures, maintenance activities, incorporation of protection technologies and protection tools; systems to detect anomalies and events by incorporating anomaly detection processes; capacity building around cyber attack response, communications processes and procedures, processes for analysis, mitigation and improvement; and recovery activities. In high-risk AI information systems, recovery plans must be in place to allow business continuity and protection against hypothetical damage to the rights of the different actors in the ecosystem in which the AI system is implemented.

Among the cybersecurity obligations added to the Annex and which makes reference to the technical documentation of Article 11(1), is pointed out that, at least, it shall contain the following information: "(…) The technical documentation referred to in Article 11(1) shall contain at least the following information, as appropriate for the relevant AI system: (…) (ga) cybersecurity measures in place".

It is worth highlighting the connection and importance given to the European Union Agency for Cybersecurity (ENISA) in order to address any emerging problems in the internal market in relation to cybersecurity, so that it will collaborate with the European Board on Artificial Intelligence.

It also establishes the obligation that "High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures", which implies that the technical and administrative measures implemented must provide for the incorporation of these elements.

2.1. Cybersecurity in high-risk AI systems deployed by authorities

When high-risk AI systems are used by authorities, they are obliged to comply with the cybersecurity rules established for these authorities. In the case of Spain, for example, the entire public sector, as defined by Article 2 of Law 40/2015 of 1 October, and in accordance with the provisions of Article 156.2 thereof. Likewise, without prejudice to the application of Law 9/1968 of 5 April, on Official Secrets and other special regulations, the systems that process classified information and the information systems of private sector entities, including the obligation to have the security policy referred to in Article 12, when, in accordance with the applicable regulations and by virtue of a contractual relationship, render services or provide solutions to public sector entities for the exercise by the latter of their competences and administrative powers, are obliged to comply with the obligations derived from Royal Decree 311/2022, of 3 May, which regulates the National Security Scheme.

Among the cybersecurity obligations to be fulfilled by entities using high-risk Artificial Intelligence systems are the organisation and implementation of the security process; risk management, consisting of a process of identification, analysis, evaluation and treatment of risks; personnel management; professionalism; access authorisation and control; protection of facilities; acquisition of security products and contracting of security services; least privilege; system integrity and updating; protection of stored and in-transit information; prevention of other interconnected information systems; logging of activity and detection of malicious code; security incidents; business continuity; and continuous improvement of the security process. Similarly, the use of common infrastructures and services of the public administrations should be integrated in order to achieve greater efficiency and feedback of the synergies of each group. It should be noted that Article 30 provides for the possibility of implementing specific compliance profiles, as well as accreditation schemes for entities implementing secure configurations and the development of capabilities that enable security auditing, security status reporting and response to security incidents.

With regard to the specific activities of prevention, detection and response to security incidents, they must comply with the technical standards, as well as with the compliance rules, which can be broken down into four: Digital Administration, service and system lifecycle, control mechanisms and procedures for determining compliance with the ENS.[30]

2.2. Cybersecurity in high-risk AI systems that are part of critical activities or essential services

High-risk AI systems also include AI systems intended to be used as safety components in the management and operation of critical digital infrastructures, rroad traffic and the supply of water, gas, heating and electricity. Similarly, the definition of serious incident includes those involving critical infrastructures, stating "any incident or malfunctioning of an AI system that directly or indirectly leads to any of the following situations (…) (b) a serious and irreversible disruption of the management or operation of critical infrastructure." In this sense, and taking into account the regulatory referrals and the integration of the set of rules, the obligations referring to critical infrastructures must be complied with.

Thus, Council Directive 2008/114 of 8 December 2008 on the identification and designation of European Critical Infrastructures and the assessment of the need to improve their protection and which in Spain is implemented

[30] Royal Decree 311/2022 of 3 May, regulating the National Security Scheme.

through Law 8/2011 of 28 April 2011, which establishes measures for the protection of critical infrastructures and Royal Decree 704/2011 of 20 May 2011, which approves the Regulation for the protection of critical infrastructures aims to establish measures for the protection of critical infrastructures, in order to specify the actions of the different bodies making up the Critical Infrastructure Protection System, incorporates a set of obligations to which the different subjects will be subject when incorporating high-risk Artificial Intelligence systems.

The cybersecurity obligations are those contained in the aforementioned provisions, highlighting, for the purposes of this document, the Operator Security Plans, which correspond to the strategic documents that define the general policies of critical operators to ensure the security of their facilities or systems, which are evaluated and approved by the Secretary of State for Security. These plans must include a risk analysis methodology that guarantees the continuity of services, addressing physical and logical threats, with criteria for the implementation of security measures, also in AI systems that are defined as high risk, as well as the mechanisms for implementation, control and monitoring.

2.3. Conformity assessments as a tool for cybersecurity in high-risk AI systems

Certifications of conformity are documents issued by certification bodies or authorised entities attesting that a product, service, system or process complies with certain previously established standards, norms or specifications. These certifications are a way of guaranteeing that a product or service complies with the requirements and standards established by the competent authorities or specialised organisations. By obtaining a conformity certification, an entity demonstrates that it has been assessed and shown to comply with the specific criteria and requirements established for its industry or sector. This may cover aspects such as quality, safety, energy efficiency, environmental sustainability, information security, among others. Conformity certifications may be mandatory for certain products or services, especially in areas regulated by government regulations. They can also be voluntary and sought by companies as a way to highlight quality and compliance with recognised standards, which can generate reliance in consumers and the market in general. Common examples of certifications include ISO 9001 certification for quality management systems, CE certification in the European Union, and various safety certifications and industry standards in different sectors.

European standardisation has as its specific legislative antecedent, among others, four different acts that need to be listed, as the cybersecurity obligations to which the new standard refers in high-risk Artificial Intelligence systems derive from them: Directive 98/34/EC of the European Parliament and of the Council of 22 June 1998 laying down a procedure for the provision of information in the field of technical standards and regulations and of rules on Information Society services[31], Decision No 1673/2006/EC of the European Parliament and of the Council of 24 October 2006 on the financing of European standardisation[32], and Decision 87/95/EEC of 22 December 1986, on standardisation in the field of information technology and telecommunications[33] and Regulation № 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC, 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council.

The new regulation on high-risk AI systems states in this respect that 'An AI system which is itself a product covered by Union harmonisation legislation listed in Annex II shall be considered as high risk if it is required to undergo a third-party conformity assessment with a view to the placing on the market or the putting into service of that product pursuant to that legislation'. (…) and that "An AI system intended to be used as a safety component of a product covered by the legislation referred to in paragraph 1 shall be considered as high risk if it is required to undergo a third party conformity assessment with a view to the placing on the market or putting into service of that product under that legislation. This provision shall apply irrespective of whether the AI system is placed on the market or put into service independently of the product".

In this regard, if high-risk AI systems or general purpose AI systems that have been certified or for which a declaration of conformity has been issued under a cybersecurity regime in accordance with Regulation (EU) 2019/881 of the European Parliament and of the Council and whose references have been published in the Official Journal of the European Union "shall be presumed to comply with the cybersecurity requirements set out in Article 15

[31] Directive 98/34/EC of the European Parliament and of the Council of 22 June 1998 laying down a procedure for the provision of information in the field of technical standards and regulations and of rules on Information Society services (OJ L 204, 21 July 1998, p. 37).

[32] Decision No 1673/2006/EC of the European Parliament and of the Council of 24 October 2006 on the financing of European standardisation (OJ L 315, 15 November 2006, p. 9).

[33] Decision 87/95/EEC of 22 December 1986 on standardisation in the field of information technology and telecommunications (OJ L 36 of 7 February 1987, p. 31).

(…) to the extent that the cybersecurity certificate or declaration of conformity or parts thereof cover those requirements".

III. Conclusions

Cybersecurity involves the development of capabilities around the process of identification, protection, detection, response and recovery. Minimising vulnerabilities and ensuring that risks do not materialise requires a process of planning and understanding of cyber security from the beginning and throughout the development process of high-risk Artificial Intelligence systems.

Effectively mitigating risks in the context of Artificial Intelligence involves imposing requirements that address various aspects, including data quality, technical documentation management, record keeping, transparency, robustness, accuracy, human oversight and, of course, cybersecurity. Although the Regulation focuses on high-risk AI systems, it is recognised that cybersecurity obligations are not exclusive to these systems, but apply to all information and communications technologies in a manner proportionate to the purposes and interests pursued.

Cybersecurity obligations are linked to high-risk AI systems contained in Annex II and III of the new regulation and those subject to third-party conformity assessment with a view to the placing on the market or putting into service of such a product.

Although cybersecurity obligations are incorporated, it is also true that the regulation directs cybersecurity to the application of existing standards and to the level appropriate to the intended purpose. Accordingly, cybersecurity standards mandated by the European Union must be applied, in particular Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity throughout the Union, amending Regulation (EU) № 910/2014 and Directive (EU) 2018/1972 and repealing Directive (EU) 2016/1148 (NIS Directive 2), and Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (European Union Agency for Cybersecurity) and the certification of information and communication technology cybersecurity and repealing Regulation (EU) № 526/2013 ("Cybersecurity Regulation"), Regulation 2022/0272, among others.

In high-risk Artificial Intelligence systems, the application of the European framework for cybersecurity certification is directed, which establishes a certification and confirmation scheme that products, processes and services associated with information and communication technologies have been evaluated and comply with requirements to protect the authenticity, integrity, availability and confidentiality of data that have been stored, transmitted or processed, or any service or function that can be accessed during the life cycle of products, services and processes.

High-risk AI systems shall comply with the obligations stemming from Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 as regards the adoption of national cybersecurity strategies, incorporating the specific section on high-risk AI systems and designating or establishing competent authorities, cybersecurity crisis management authorities, cybersecurity single points of contact and cybersecurity incident response teams.

Finally, the linkage and importance given to the European Union Agency for Cybersecurity (ENISA) in order to address any emerging problems in the internal market in relation to cybersecurity should be highlighted, so that it will collaborate with the European Committee on Artificial Intelligence.

POST-MARKET MONITORING ON HIGH-RISK AI SYSTEMS IN THE ARTIFICIAL INTELLIGENCE ACT. DESCRIPTION, MEASURES AND USE CASES

Idoia Salazar

PhD. Lecturer at CEU San Pablo University. President of Odiseia

Miguel Ángel Liébanas

Criminologist expert in Intelligent Systems. Odiseia. CEO of Human Trends

I. Post-market monitoring in the Regulation

1. Introduction

The rapid evolution and adoption of Artificial Intelligence (AI) systems in diverse areas, from medicine to national security, have brought with them a number of significant benefits. However, the inherent complexity and evolving capabilities of these systems pose unique challenges in terms of security, privacy, ethics and governance. Specifically, high-risk AI systems, those whose malfunction or misuse could have serious consequences for individuals or society, require special consideration. In this context, post-market surveillance emerges as a critical component to ensure that these systems operate safely, effectively and ethically throughout their lifecycle.

This Article 72 AIA concerns "Post-market monitoring by providers and post-market monitoring plan for high-risk AI systems". It highlights the importance of implementing a robust and systematic post-market surveillance plan for these high-risk AI systems. It shows that such plans are essential to identify and mitigate emerging risks associated with the long-term use of AI. Furthermore, their crucial role in building public trust and promoting accountability and transparency on the part of AI developers and users is highlighted. In this regard, key issues such as identifying risk indicators, continuous performance monitoring, managing user feedback and adapting to changing technological and social dynamics are addressed. On the other hand, it also highlights the implications of insufficient post-market surveillance, including the risks of unwitting harm, loss of public trust, and potential regulatory barriers that could inhibit responsible innovation.

In any case, it is intended to underline that post-market monitoring is not simply a regulatory obligation, but a strategic opportunity for AI developers and deployers. It is a way to ensure that AI systems - whether high-risk or not - not only meet their initial objectives, but must also adapt and improve responsibly in response to emerging challenges and societal expectations.

2. What is a post-market monitoring plan and what does it include?

Article 3.25 defines a "post-market monitoring system" as "s all activities carried out by providers of AI systems to collect and review experience gained from the use of AI systems they place on the market or put into service for the purpose of identifying any need to immediately apply any necessary corrective or preventive actions". Thus, a Post-Market Monitoring Plan is a set of processes and tools aimed at collecting data from a system and transforming it into a set of indicators on its activity with the objective of monitoring Artificial Intelligence (AI) systems after they have been introduced on the market. In this sense, such a plan would include a series of tasks, always taking into account the intended purpose of the system. These would be the following:

- Collecting data on system performance and security.
- Assessment of the possible causes of identified problems.
- Implementation of solutions to correct problems.
- Communication of results and recommendations to stakeholders.

Each of these is described in more detail below:

A) Collecting system performance and security data

This involves proactively monitoring the AI system to gather information about its operability and any adverse events or deviations from expected behaviour. Data can come from a variety of sources, including error logs, user *feedback* and other anomaly detection systems. Several components are important to consider in this regard:

- *Defining relevant metrics:* Before collecting data, it is critical to define which metrics accurately reflect the performance and security of the system. These may include prediction accuracy, processing speed, failure rate, frequency of false positives or negatives, and other indicators of stability and reliability. These are discussed in more depth in the Indicators section.

- *Real-time monitoring systems:* Implement systems that continuously monitor the AI system for anomalies. These systems should be able to record events in real time and provide early warnings of potential security or performance issues.

- *Gathering feedback data:* System users are often a rich source of feedback. They can report problems that are not obvious to automated monitoring systems, thus providing a more holistic view of the performance and safety of the AI system.

- *Impact analysis:* Beyond technical data collection, it is important to understand the impact of AI on the environment in which it is deployed. This may include analysing how AI decisions affect people, business processes and other technology systems.

- *Secure data sharing:* Given that this data may include sensitive information, it is crucial to ensure that the collection and storage of this data is done in a secure manner and in compliance with applicable data privacy regulations.

- *Benchmarks and stress testing:* On a regular basis, the AI system should be performance tested against benchmarks or industry standards to evaluate its performance under different conditions and workloads.

- *Recording and documentation:* Keeping a complete record and detailed documentation of all data collected is important for long-term monitoring and auditing of the system. This also helps to establish a baseline for understanding its evolution over time.

- *Incorporation of new data:* AI systems can deviate from their expected performance as they are exposed to situations not anticipated in their training phase. Incorporating new data collected during operation into AI training can help the system adapt and improve over time.

- *Continuous assessment of data adequacy:* As the environment and contexts change it must be assessed whether the data on which the AI is based is still representative and adequate for the task it is supposed to perform.

B) Assessment of possible causes of identified problems

When problems within a system are identified, a deep analysis phase proceeds to decipher the root causes behind these mishaps. This analysis may require the use of advanced data management methodologies, such as machine learning and data mining, to reveal patterns and correlations that are not immediately obvious.

In this regard, the first step involves meticulous scrutiny of the incident, which may include reviewing logs, identifying the specific conditions under which the problem manifested itself, and interacting with impacted users to understand their experience. Diagnostic tools are then used to examine the internal state of the AI system, including the review of debug logs and the use of performance monitors, in order to understand how the system operates.

On the other hand, the application of causal analysis is essential to determine the connections between various factors and the identified problem by analysing the sequence of events that led to the incident and examining both the input data and the decisions made by the AI system. To verify hypotheses about potential causes, controlled experiments or simulations are conducted to assess whether the problem can be reproduced under the same conditions.

It is very important to review the source code and algorithms used by the AI system to detect possible bugs or errors in the logic that could be causing the problems. In addition, data mining and machine learning techniques are used to discover hidden patterns that may be contributing to the problematic situation.

As we have seen above, reviewing the data used to train the AI system is another critical aspect. This ensures that it is complete, accurate and unbiased, as deficiencies in the training data can translate into inadequate system performance. End-user feedback provides valuable insights into how the system behaves in real-world environments and how these behaviours are related to the detected problems.

Equally important in this area is the assessment of the operating environment in which the AI is deployed, as external factors, such as changes in hardware, complementary software or environmental conditions, can influence the performance of the system.

Finally, consultation with domain experts such as software engineers, data scientists or specialists in the AI application domain can also provide additional insights into the causes of problems.

C) Implementation of solutions to correct problems

Based on the assessment of the causes, solutions are designed and implemented. This may include updating algorithms, modifying data sets, revising automatic decision processes or improving security protocols.

Agility and efficiency in this process is important to mitigate risks and prevent escalation of problems. In this regard, it is necessary to consider prioritisation of issues, development of fixes, rigorous testing, and review ethical and regulatory implications. Before briefly going into each point, it is important to stress that it is preferable to implement the solution gradually, first in a test environment, then to a small group of real users, and finally to the entire user base, to minimise risk.

- Regarding Problem Prioritisation: Based on the severity and impact of the problems identified, a priority is set to address them. This involves considering the risk to users and the organisation, as well as the frequency and consequences of the problem. Since the risk analysis of the smart system will have been developed previously, we can use this as a model to determine prioritisation.

- Regarding the development of fixes: Work to develop specific fixes, either in the code, in the algorithms, or in the data used by the AI. In some cases, this may mean retraining models with new data or adjusting model parameters.

- Regarding rigorous testing: Before any solution is implemented, extensive testing is performed to ensure that it not only solves the problem but also does not introduce new problems. This may include unit testing, integration testing, system testing, and user acceptance testing.

- Regarding the review of Ethical and Regulatory implications: Each proposed solution should be reviewed to ensure that it complies with applicable regulations (AIA or other applicable regulations depending on the country where the solution is implemented) and adheres to ethical standards, especially in terms of privacy and fairness.

D) Communication of findings and recommendations to stakeholders

It is vital to maintain an open and transparent dialogue with all stakeholders, including regulators, end-users, and the general public. Effective communication about how problems are being handled and improvements implemented is essential to maintain trust in the AI system.

These tasks are integrated into a governance and risk management framework that must also include adaptability and continuous improvement. As the AI system learns and evolves, so does the understanding of its potential risks, requiring a dynamic approach to post-market surveillance management.

3. Why is a post-market monitoring plan necessary?

The creation of a post-market surveillance plan for high-risk AI systems is not only a necessity under the AIA, but is also highly recommended to maintain the efficiency and security of the system over time. In addition, this type of monitoring is key to maintaining public confidence and for systems to adapt to changing needs and data.

This ‘continuity’ element in the security and efficiency of AI systems is crucial; without proper monitoring, they may face unexpected problems or a decrease in performance due to changes in data patterns or in their operating environment. Post-market monitoring enables early detection and correction of these problems, preventing significant damage and preserving reliance in the technology.

Moreover, adapting to new data and contexts is important in today’s dynamic technological environment. AI systems, especially those based on machine learning, require regular updates to remain relevant and effective. An effective monitoring system ensures that these updates are made in a timely manner, allowing AI to respond appropriately to new and unforeseen situations. On the other hand, responsible innovation is a key objective in this plan. By monitoring the performance and impacts of AI systems after their release, developers can identify areas for improvements and technological breakthroughs. This not only prevents risks but also fosters ethical and sustainable innovation.

Regulatory compliance, specifically AIA, is also an important consideration, as AI regulations are constantly evolving. A post-market monitoring plan ensures that AI systems remain compliant with current regulations, avoiding legal sanctions and protecting users.

We conclude this section by again highlighting the importance of user trust and transparency. An oversight system that promotes accountability can strengthen public trust in AI by demonstrating a continued commitment to safety and accountability.

4. Post-market monitoring plan in the Regulation

Chapter IX of the AIA specifies the rules on post-market monitoring, information sharing and market surveillance. In particular, Section 1 refers exclusively to post-market monitoring.

Section 1 has only Article 72, which focuses on post-market monitoring by providers and post-market monitoring plan for high-risk AI systems.

First, paragraph 72.1 indicates the obligation to establish and document a post-market monitoring system in a manner that is proportionate to the nature of the deployed intelligent system.

> "Providers shall establish and document a post-market monitoring system in a manner that is proportionate to the nature of the AI technologies and the risks of the high-risk AI system." (Art. 72.1, Chapter IX, AIA).

The following section provides an overview of the functions of such a monitoring system and its purpose: to assess that the requirements of Chapter IX, Section 1, are maintained throughout the lifecycle of the intelligent system.

> "The post-market monitoring system shall actively and systematically collect, document and analyse relevant data which may be provided by deployers or which may be collected through other sources on the performance of high-risk AI systems throughout their lifetime, and which allow the provider to evaluate the continuous compliance of AI systems with the requirements set out in Chapter III, Section 2. Where relevant, post-market monitoring shall include an analysis of the interaction with other AI systems. This obligation shall not cover sensitive operational data of deployers which are law-enforcement authorities" (Art. 72.2, Chapter IX, AIA).

The third paragraph of Article 72 focuses on how to develop the monitoring system based on the design of a post-market monitoring plan. In section II "Addressing the Post-Market Monitoring Requirement" of this analysis this point will be addressed in more detail.

Finally, paragraph 4 sets out a number of exemptions where it is not necessary to develop a post-market surveillance system:

High-risk AI systems covered by the Union harmonisation legislation listed in Section A of Annex I, where a post-market monitoring system and plan are already established under that legislation, with an equivalent level of protection. For example, high-risk systems related to the safety of toys.

It also applies to high-risk AI systems referred to in point 5, Annex III placed on the market or put into service by financial institutions.

Thus, the AIA establishes the need for the post-market monitoring system, states that its objective is to maintain the Title III, Chapter 2 intelligent system control criteria throughout the lifecycle of the system, and determines that the monitoring plan will be the basis on which to design and implement the system.

5. Who should conduct the post-market monitoring system?

The implementation and management of a post-market monitoring system, as stipulated in the AIA regulatory framework, rests primarily with the providers of AI systems. This approach ensures that AI systems, once implemented and operational, continue to comply with established standards and regulations throughout their lifecycle. It is imperative that providers take responsibility for assessing and ensuring the ongoing compliance of their systems with relevant legal and ethical requirements, including aspects of security, privacy, and transparency.

In this regard, providers should establish robust procedures for the ongoing monitoring of their AI systems. This involves not only a technical review of the system performance but also consideration of how changes in the operating environment or data may affect AI efficiency.

In addition, providers have a responsibility to design mechanisms for *feedback* collection and analysis*,* including the reporting of incidents and anomalous behaviour by users. This means that AI systems must be designed with capabilities to record and report any failures or deviations in their behaviour, thus facilitating an efficient feedback process between deployers and providers.

User responsibility.

On the other hand, (non-end) users of AI systems also play an important role in this ecosystem by acting as active observers of the technology in use. Deployers are expected to report any incidents, failures or unusual system behaviour to the vendor. This collaboration between deployers and providers is essential for early detection of problems and to ensure that corrective action is taken in a timely manner.

The synergy between providers and deployers, supported by a clear regulatory framework such as the AIA, facilitates an environment where AI systems are not only constantly monitored and evaluated but where continuous improvement is encouraged. This ensures that AI systems maintain high levels of reliability, safety, and regulatory compliance while adapting to society's changing needs and technological advances.

II. Addressing the Post-Market Monitoring requirement

After presenting the concept of post-market surveillance, we will understand how to approach this requirement from a procedural and technical perspective. In the following sections, some technical concepts for its implementation will be presented, but always accompanied by a brief explanation that will facilitate its understanding without requiring any prior knowledge.

1. Monitoring Plan and Monitoring System. Key elements of monitoring

The main objective of Post-Market Monitoring is to verify that the requirements set out in Chapter III of the AIA are met throughout the lifecycle of the intelligent system. To achieve this objective, we need two interrelated components.

First, it is necessary to have a set of processes and protocols with which to define the surveillance activity *per se*. On the other hand, a monitoring system will be required from a technical perspective to obtain all the necessary metrics from the intelligent system, process them in due time for analysis, and, if necessary, obtain the corresponding alerts in case of incidents. These two systems can be defined as follows:

- *Post-Marketing Monitoring Plan*. Although the AIA does not provide a definition of the Monitoring Plan, it can be inferred from the articles that it is a set of protocols and designs that give structure and operability to the Post-Marketing Monitoring System. In other words, it is the plan that must be followed to achieve satisfactory monitoring of the system. Within this plan, the tasks to be performed, the responsibilities of the personnel associated with the system, the technical design of the monitoring system itself, and the documentation of all its contents in the technical documentation of the intelligent system must be established. Therefore, the design of the Monitoring System will be part of the content of the Monitoring Plan:

> *"The post-market monitoring system shall be based on a post-market monitoring plan. The post-market monitoring plan shall be part of the technical documentation referred to in Annex IV"* (Art 72.3, Chapter IX, AIA).

- With regard to the content of the Plan, the AIA itself indicates that:

> *"T. The Commission shall adopt an implementing act laying down detailed provisions establishing a template for the post-market monitoring plan and the list of elements to be included in the plan by 2 February 2026"* (Art. 61.3, Chapter IX, AIA).

Therefore, we do not yet have the template and requirements for a post-market monitoring plan. However, this analysis will indicate the minimum content of any monitoring plan that is likely to be part of the requirements set by the Commission.

> - *Post-market monitoring system.* The AIA indicates that the Monitoring System are: "*all activities carried out by providers of AI systems to collect and review experience gained from the use of AI systems they place on the market or put into service for the purpose of identifying any need to immediately apply any necessary corrective or preventive actions;*" (Art. 3.25, Chapter I, AIA).

Although the text refers to activities, it is more intuitive to think of the system as a set of automated (and exceptionally manual) processes that aim to obtain the necessary data to assess that the requirements of Chapter III as applied to the intelligent system are still being met. Also, this system should be able to detect any failure in the system as early as possible, always acting preventively if possible.

In this way, the Monitoring Plan will address the procedures necessary to develop the monitoring of the smart system, and the Monitoring System will be the tool that will allow this work to be carried out. The Plan will define what, when, how, and by whom the monitoring will be carried out, and the System will be the tool to support all this work.

2. Design of the Monitoring System

The aim of this section is not to go into the technical requirements and useful tools for developing the Monitoring System but to get an intuition of what components make up such a system and what requirements we should expect from the system.

To get an overall view of the system, it is best to start at the end, i.e., the outputs of the system. In this case, these are the system's monitoring panel and the preconfigured alert system.

First, the monitoring panel will allow market surveillance authorities to monitor that all indicators (a concept that will be discussed in the next section) remain in their normal range and that the system is functioning as expected.

Secondly, in order to maintain continuous monitoring and to be able to respond quickly to incidents and even prevent them, an alert system with the necessary communication protocols is needed to keep those in charge informed of any changes in the system's indicators.

It will now be defined where the monitoring panel and the alert system will get the data from to assess the current status of the system. For this purpose, a database system will be used in which all records of the selected indicators will be stored. The type of database will depend on the architecture of the intelligent system itself and the use case developed. Ultimately, this storage system will function as the hub of information from the activity logs of the deployed smart systems and will provide this information to the monitoring panel and the alert system.

Finally, a system for sending logs from the deployment point of the intelligent system to the log storage system shall be implemented. As an example, if the intelligent system performs its processing on the device itself, such as a security camera with intelligent anomaly detection, a system for sending logs from the devices to the storage system shall be implemented with a periodicity determined in the design of the monitoring system.

In this way, we can divide the architecture of the Monitoring System into three blocks:

- *Log sending system*: It will send the system indicators from the device where the processing is carried out. This device can be a server, an IoT system, or any system where the processing of the smart system takes place.

- *Log storage system*: shall store all indicators generated by the intelligent systems.

- *Monitoring Panel and Alert System*: will process the information received, providing actionability and accessibility to those responsible for monitoring.

3. The Indicator concept

The term indicator has been mentioned previously, but its meaning and practical application in the monitoring system has not been elaborated on. What is the importance of indicators in this work? Let us look at a practical example.

If we are told to monitor the temperature of a car because it has been malfunctioning for the last few days, what would we ask? We would probably ask them to tell us what the normal temperature is and at what temperature it is considered to be a problem, both high and low temperature. With that data, we would then have a scale, and we could monitor that the temperature of the car remained stable.

The concept of an indicator is exactly the above. We can define it as a piece of data framed within a scale that allows us to infer that we are approaching a specific scenario or that we are leaving the scenario we are trying to measure. In our example, the data is the temperature, the scale is the limits that the mechanic has told us and the scenarios are the normal operation of the vehicle or the faulty state of the vehicle.

In the context of intelligent systems, indicators will be based on data about the performance of the system, for example, in the case of the surveillance camera, the number of predictions or images processed per minute. For each data, we will have to establish its scale of normality and define which scenarios we are trying to monitor. For example, a drastic reduction in the number of images processed could indicate an overload of the device and lead to a malfunction.

There is no concrete list of indicators for our smart systems, but we should create one together with the technical team to try to monitor how close we are to the risks detected in the risk analysis developed and to a general malfunctioning of the system.

Of course, not all logs can be considered as indicators. There are logs that will indicate previously unknown system errors in text format that will need to be evaluated by the technical team. However, it is vital to have as complete an outline of indicators as possible from the perspective of the intelligent system, the device that supports it, the end-user usage and the cyber security risks highlighted.

Finally, the monitoring and warning system will use these indicators as a basis for making the whole monitoring system actionable as a whole.

4. Measures to be developed in the Monitoring Plan

We do not yet have a model or requirements for a post-market monitoring plan. That said, there are certain basic pillars that should be introduced in any monitoring plan and which will most likely be requirements in the model offered by the Commission. Specifically:

- *Continuous Monitoring*: This is monitoring with a very reduced periodicity in time (minutes or hours depending on the system) that will allow those responsible to detect abrupt changes in the operation of the intelligent system. The alert system will allow this work to be carried out reactively in the shortest possible response time if an anomalous scenario occurs. However, in many cases, this early warning system should not be the only monitoring task and a continuous verification that all indicators remain stable should be developed.

- *Periodic monitoring*: In this case, instead of real-time monitoring, this monitoring is carried out on a more extended basis (days, weeks, or months) in order to evaluate palliative and time-delayed changes that may go unnoticed in real-time monitoring. For example, if we evaluate the accuracy of a smart system hour by hour, we may not see any change, but if we develop a weekly evaluation, we can see if there has been a significant change.

- *Assignment of responsible persons*: The Plan should assign the persons responsible for carrying out the monitoring activities and the maintenance of the monitoring system.

- *Training*: The persons selected as responsible persons should know about the functioning of the monitoring panel, the alert system, the indicators of the system and the established reporting protocols.

- *Incident reporting protocol*: It is of vital importance to develop a protocol for reporting and recording incidents detected in the system. In the case of serious incidents, this protocol should be complemented by the measures established in the AIA for these cases, such as notification to the Surveillance Authority.

As mentioned above, the official model and requirements of the Monitoring Plan are not yet known, but it is highly likely that the points explained above will be among those selected.

5. Validity of the Monitoring System and Plan

Once the Monitoring System and the Monitoring Plan have been developed, the question arises as to how long this system will be valid for the correct monitoring of the intelligent system. The answer is straightforward: as long as the intelligent system does not undergo any modifications that alter the Monitoring Plan or System or a new risk analysis is carried out that exposes new risk scenarios that must have new indicators.

III. Conclusions

Post-market monitoring is an indispensable component in the development and deployment of high-risk AI systems, as it ensures that these systems are safe, effective, and ethically responsible throughout their lifecycle. This process requires close collaboration between providers, deployers, and regulators and should be seen as an opportunity to continuously improve AI technology, promote its social acceptance, and foster responsible innovation in the field. Thus, as a result of the analysis conducted, we detail the following conclusions:

Post-market monitoring is essential to monitor and maintain the safety, efficacy, and ethical and regulatory compliance of high-risk AI systems throughout their lifecycle. This ongoing process helps identify and mitigate emerging issues that may not be evident in the design and testing phases.

The primary responsibility for implementing robust monitoring systems lies with the providers of high-risk AI systems. This includes constantly monitoring system performance, adapting to changes in the operating environment, and responding to ethical and legal challenges that arise during system use. On the other hand, high-risk AI systems operators also play an important role in post-market monitoring. They provide feedback on system performance and report any incidents or anomalies. This collaboration is critical for early detection of problems and timely implementation of solutions.

One of the biggest challenges in post-market monitoring is the ability of AI systems to adapt to constantly changing operating environments and evolving data sets. This requires flexible and dynamic monitoring mechanisms that can adjust to new conditions and challenges.

Post-market monitoring plays a crucial role in ensuring that AI systems continuously comply with evolving regulations and ethical standards. This issue not only protects users and society but also ensures trust and acceptance of AI.

In conclusion, post-market monitoring is an indispensable component in the development and deployment of high-risk AI systems, ensuring that these systems are safe, effective, and ethically responsible throughout their lifecycle. This process requires close collaboration between providers, deployers and regulators and should be seen as an opportunity to continuously improve AI technology, promote its social acceptance, and foster responsible innovation in the field.

General-purpose Artificial Intelligence, non-high-risk systems and Article 50 systems

GENERAL-PURPOSE ARTIFICIAL INTELLIGENCE, FOUNDATIONAL MODELS (AND "GPT CHAT") IN THE ARTIFICIAL INTELLIGENCE ACT

José Antonio Castillo Parrilla[1]
PhD. Ramón y Cajal Researcher - University of Granada

I. Introduction

AI is fully embedded in everyday life[2]: virtual diary assistants, translators, video subtitle generators, tools on platforms for content suggestions, and a long etcetera of examples. The European Commission states that AI refers to systems that exhibit intelligent behaviour by analysing their environment and taking actions (with a certain degree of autonomy) to achieve specific objectives[3]. AI-based systems can be purely software-based (e.g., recommender systems or search engines), or integrated in hardware (robots, drones, or IoT applications). In 2019, the European Commission's High Level Expert Group on AI defined AI as software (or hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment, acquiring data, interpreting the collected data, whether structured or unstructured, reasoning about the knowledge, processing the information derived from this data, and deciding the best action(s) to take to achieve the given goal.[4]

The relevance of general-purpose AI models and systems has been growing considerably in recent years, especially since the emergence of Chat

[1] This publication is part of the contract RYC2021-031430-I, funded by MCIN/AEI/10.13039/501100011033 and by the European Union, "NextGenerationEU". It is carried out in the framework of the GOIA Project, funded by MCIN/AEI/TED2021-12902B-C22 and by the European Union, "NextGenerationEU". Much of the research was carried out during the stay at the Universidade Nova de Lisboa from March to May 2024, funded by the Plan Propio of the University of Granada.

[2] https://ec.europa.eu/commission/presscorner/detail/en/statement_23_6474

[3] Commission, Communication "Artificial Intelligence for Europe", Brussels, 25 April 2018, COM (2018) 237 final, https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe, p. 1.

[4] HLEG-AI - High Level Expert Group on Artificial Intelligence, (2019): A definition of AI: main capabilities and disciplines, European Commission, Brussels, April, https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

GPT at the end of 2022, which has caused a real earthquake[5]. Chat GPT is a text-generative AI tool (*large language model* or LLM) that was launched on 30 November 2022[6], initially for free, and in just five days surpassed one million users and 180 million active users by November of the following year[7]. On 14 March 2024 OpenAI launched GPT-4[8], which in less than six months surpassed 100 million weekly active users[9]. Chat GPT is not the only generative AI tool[10], but it is one of the most popular today, and has sparked intense debate about the risks posed by this type of feature; So much so that, for example, the European Data Protection Committee initiated in April 2023 a working group on Chat GPT[11] following the decision of some national data protection authorities such as the AEPD[12] or the Garante Privacy (Italy)[13] to initiate ex officio investigation proceedings[14] for possible breach of data protection regulations.

The social impact and popularity of this tool has been such that if the 2021 AIA Proposal[15] did not mention foundational or general-purpose AI models, barely three months after the launch of GPT-4, the Amendments to the text presented by the European Parliament[16] dedicated several new

[5] Novelli, C. et al., "Generative AI in EU Law: Liability, Privacy, Intellectual Property and Cybersecurity", Cornell University, https://arxiv.org/abs/2401.07348.

[6] https://openai.com/blog/chatgpt, p. 1.

[7] https://www.primeweb.com.mx/chatgpt-usuarios-estadisticas Among the data highlighted, it is worth noting that nearly 80% of young people between 18 and 29 years old have used or have seen someone else use the tool.

[8] https://openai.com/research/gpt-4

[9] https://www.theverge.com/2023/11/6/23948386/chatgpt-active-user-count-openai-developer-conference

[10] Other generative AI tools have been developed in the field of so-called "Artificial Intelligence art", such as Stable Difussion, Midjourney or DALL-E. In generative AI of text Chat GPT is also not unique: Microsoft has launched Copilot, and Nvidia RTX.

[11] https://www.edpb.europa.eu/news/news/2023/edpb-resolves-dispute-transfers-meta-and-creates-task-force-chat-gpt_en

[12] https://www.aepd.es/prensa-y-comunicacion/notas-de-prensa/aepd-inicia-de-oficio-actuaciones-de-investigacion-a-openai

[13] https://www.garanteprivacy.it/home/docweb/-/docweb-display/docweb/9870847

[14] The National Commission for Data Protection (CNPD, Portugal) only expressed its interest in this issue but did not initiate an ex-officio investigation on Chat GPT: https://observador.pt/2023/04/03/chatgpt-cnpd-leu-com-muito-interesse-decisao-de-bloqueio-em-italia-mas-nao-preve-para-ja-algo-semelhante-em-portugal/

[15] https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A52021PC0206

[16] European Parliament (2023): Amendments adopted on the Proposal for a Regulation laying down harmonised rules in the field of Artificial Intelligence (COM (2021) 0206), https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_ES.html.

Recitals and articles to foundational models; and the text finally presented and approved as AIA[17] dedicates an entire Chapter to the regulation of general-purpose AI models, in addition to various specific mentions and obligations in other sections of the AIA.

It is also indicative of the importance of the issue that the inclusion (and scope) of general-purpose AI in the AIA was one of the last aspects of discussion in the final negotiation of the text. In November 2023, there were public debates between various Member States on the then so-called foundational models during the trilogue phase, despite a consensus on the need to include certain transparency rules: Germany, France, or Italy were in favour of promoting the development of codes of conduct but without a regime of sanctions already set by the AIA[18], while Spain advocated the inclusion of obligations beyond transparency, and even to address the challenge of copyright[19].

II. General Artificial Intelligence, general-purpose Artificial Intelligence, foundational models and generative Artificial Intelligence

The first major classification into which AI tools are divided is that which distinguishes those based on logical rules from those based on data[20]. The former, also known as expert systems, are capable of carrying out tasks very well in delimited and relatively simple fields based on the incorporation of logical rules and the knowledge of experts (who, again, design logical rules that the machine incorporates). The most popular example today is perhaps still Deep Blue[21]. Data-driven AI tools are fed with large amounts of data that allow them, through various techniques (*machine learning*, neural networks, *deep learning*, decision trees…) to solve unspecific problems or problems whose solution cannot be reached through deductive reasoning (text or image analysis, behaviour prediction, or recommender systems[22]). Although data-driven

[17] https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_ES.pdf

[18] https://www.euractiv.com/section/artificial-intelligence/news/france-germany-italy-push-for-mandatory-self-regulation-for-foundation-models-in-eus-ai-law/

[19] https://www.euractiv.com/section/artificial-intelligence/interview/eu-ai-act-cannot-turn-away-from-foundation-models-spains-state-secretary-says/

[20] Valls Prieto, J., Inteligencia artificial, derechos humanos y bienes jurídicos, Aranzadi, Navarra, 2021, p. 20.

[21] IBM's Deep Blue tool defeated the then world chess champion Gary Kasparov in 1997 in the second match (6-game match) played that year, following a first match in 1996 from which Deep Blue "learned" (https://www.ibm.com/history/deep-blue).

[22] Recommender systems (e.g., used in online marketing platforms or techniques in social

AI systems always have a certain error rate, this will be reduced as they are able to obtain and process larger amounts of data[23].

In the field of AI, a distinction is often made between narrow or weak AI and general or strong AI. General or strong AI is defined as an AI system capable of performing typically human activities, while narrow or weak AI is defined as an AI system capable of performing one or a few specific tasks. While most of the AI systems developed up to 2019 could be qualified as narrow or weak AI[24], recent developments and their popularisation seem to justify the AIA's decision to regulate what it calls general-purpose AI. The term general-purpose AI should therefore not be identified with the terms used by the AIA (general-purpose AI model / general-purpose AI system), but merely as "strong AI".

There have been some nuances in the texts between June 2023 and March 2024 that need to be clarified at this point. The most relevant change in terminology, however, is the replacement of "foundational models" by "general-purpose AI model". These changes in terminology should be borne in mind not only to analyse how much of the European Parliament's proposals have been integrated into the text of the AIA, but also, as far as Spain is concerned, because RD 817/2023 of 8 November followed the European Parliament's terminology and therefore speaks of foundational models and general-purpose AI systems[25].

To avoid unnecessary complications, hereafter we will simply speak of general-purpose AI models following the AIA terminology. General-purpose AI models are AI models (which may be trained on a large amount of data using large-scale self-monitoring), which exhibit a considerable degree of generality and are capable of competently performing a wide variety of different tasks, as well as being integrated into a variety of downstream systems or applications (Art. 3.63 AIA).

Finally, generative AI is a type of AI based on general-purpose AI models that can flexibly generate text, audio, image, or video content. It is therefore a

media and search engines) are based on a technique called *reinforcement learning*: the AI system is allowed to make decisions freely, and is rewarded when it gets it right. The goal of the AI system is to maximise rewards.

[23] HLEG-AI, op. cit., pp. 3-4.

[24] HLEG-AI, op. cit., p. 5.

[25] Royal Decree 817/2023 of 8 November establishing a controlled test environment for testing compliance with the proposal for a Regulation of the European Parliament and of the Council laying down harmonised standards in the field of Artificial Intelligence, cf. articles 3.6 and 3.5. Available at: https://www.boe.es/diario_boe/txt.php?id=BOE-A-2023-22767

specific category of general-purpose AI models[26]. Generative AI raises issues related to the generation of new content and the respect of intellectual property rights[27]: generative AI models follow the same functioning as general AI models, i.e., they are fed by a large amount of input information, most of which may be protected by intellectual property rules, with the consequent need to seek consent where appropriate (Recital 105 AIA).

III. What is and what is not general-purpose Artificial Intelligence in the Regulation? general-purpose models with systemic risk and exclusion of models specifically intended for research purposes

Although AI models are essential components of AI systems, they do not constitute systems by themselves, as they need other components (e.g., a user interface) to become AI systems (Recital 97). Therefore, between models and AI systems there is a whole/part relationship, where the AI system is the whole and AI models can be a part. Within data-driven AI systems (i.e., those that in order to achieve their objectives use inferences based on input information to produce various products as output information *ex Art.* 3.1 AIA), general-purpose AI systems are considered to be those that are based on a general-purpose AI model (Art. 3.66 AIA).

Two fundamental characteristics define an AI model as general-purpose: (1) generality, and (2) the ability to competently perform a wide variety of differentiated tasks. What is generality to be understood? The AIA does not define generality in Art. 3, but it does provide a benchmark: models are general if they: (1) have at least one billion parameters, and (2) have been trained on a large volume of data using large-scale self-monitoring (Recital 97).

Generative AI models are a sub-category within general-purpose AI models (Recital 99). general-purpose AI models with systemic risks can also be understood as a sub-category (Recital 111 and Art 51 AIA).

A general-purpose AI system should be considered to present systemic risks when: (1) it has high impact capabilities, or (2) it has a significant impact on the internal market due to its scope. The second possibility does not seem to include content aimed at disinformation, as this is mainly limited to mere transparency obligations[28]. High-impact capabilities are those that match or exceed those shown by state-of-the-art general-purpose AI models, which

[26] Recital 99 AIA, and Recital 111 and 51 AIA in respect of general-purpose AI models with systemic risk.

[27] European Parliament Amendment 102.

[28] Watcher et al., op. cit., p. 41.

will depend on the state of the art at the time. An important benchmark for determining whether a general-purpose AI model has high-impact capabilities is the FLOPS threshold, i.e., the cumulative amount of computations used for training the general-purpose AI model, measured in floating point operations (Art. 3.67 AIA). This threshold should be adjusted in line with technological developments to reflect technological and industrial changes, algorithmic improvements or increased hardware efficiency (Recital 111).

Annex XIII of the AIA sets out benchmark criteria to be taken into account when assessing whether a general-purpose AI model has high-impact capabilities (Art. 51.1.b AIA). These criteria are: (1) the number of model parameters; (2) the quality and size of the dataset; (3) the amount of computation used to train the model, measured in FLOP or in combination with other variables; (4) the input and output modalities of the model (text to text, text to image…); (5) the benchmarks and assessments of the model's capabilities; (6) the significance of its impact due to its scope; and (7) the number of registered end-users.

The criteria are open-ended: an exemplary list is established, and for each of the criteria no minimum thresholds are introduced, except for two: the FLOPS threshold (in Art. 51.2 AIA), and the scope of the model, which will be assumed when it has been made available to at least 10,000 registered professional users established in the EU[29]. This system is consistent with the fact that the Commission will designate a generally used AI model as a model with systemic risk (Art. 52.1 AIA). Moreover, the rule provides for its own capacity to adapt in a flexible manner in Art. 52.3 AIA, which empowers the Commission to adopt delegated acts to amend the criteria and thresholds referred to in the light of technological developments (Art. 97 AIA and 290 TFEU). This is in the interest of legal certainty for providers of AI systems, and particularly in this respect for providers of general-purpose AI models[30].

Legal certainty is also reinforced in the procedure for the consideration of a general-purpose AI model as a model with systemic risk (Art. 52 AIA).

[29] Annex XIII, point f. However, if we look at the wording of this paragraph, we can see that it is also an open criterion: the scope is deemed to be exceeded when the 10 000 registered professional users established in the EU are reached; but nothing prevents the model from being considered to have a significant scope with a lower figure, especially bearing in mind that this is only one of the criteria to be taken into account by the Commission when designating an AI model for general use as a model with systemic risk.

[30] It is also in the interest of legal certainty and transparency for the Commission to publish an updated list of general-purpose AI models with systemic risk, always providing sufficient information on them but without jeopardising intellectual property rights, industrial property rights and trade secrets of the models (art. 52.6 AIA).

This is arguably a "two-layered" procedure: (1) proactive responsibility of the provider; and (2) review and closure proceedings by the Commission. The procedure is appropriate insofar as it not only favours legal certainty in the traffic of these products, but also places an important part of it on those who benefit economically from the models, leaving the institutions (the Commission in this case) with a role of control and closure of the system.

The first phase (art. 52.1 AIA) relies primarily on the provider of the general-purpose AI model, which must notify the Commission that it has passed or intends to pass the requirements that qualify it as a model with systemic risk, providing the necessary documentation. It may also provide, together with this documentation, arguments that would support the absence of systemic risk in the specific case due to the specific characteristics of the AI model (art. 52.2 AIA), which the Commission will assess for the purposes of, finally, its designation or not as a general-purpose AI model with systemic risk (art. 52.3 AIA)[31]. The provision should be interpreted as meaning that the provider is obliged to notify the Commission within two weeks from the day after it has met a certain easily verifiable requirement: exceeding the FLOP threshold or the number of users in Annex XIII.f in force at any given time, or other similarly clear thresholds that may be established in the future[32]. Once this deadline has passed, we would enter the second phase, in which the Commission can review and, if necessary, complete the task of proactive liability of providers of general-purpose AI models.

[31] This process may be repeated in the future (not earlier than 6 months after designation) if the provider of the model already designated as a general-purpose AI model with systemic risk requests reassessment by the Commission, provided that it provides new reasons since designation that justify a change (Art. 52.5 AIA).

[32] The wording of Art. 52.1 AIA could be improved for several reasons. Firstly, Art. 51.1.a AIA does not speak of "a requirement", but of several, most of which are not associated with objective thresholds except for two: FLOP and scope of the model measured in number of active professional users registered in the EU. It should therefore be understood that when a general-purpose AI model exceeds the FLOP threshold of Art. 51.2 or the threshold of Annex XIII.f, or those that may replace them in the future, it must notify the Commission of this circumstance. The second aspect of improved clarity in the wording of Art. 52.1 concerns timing: both in the determination of the dies a quo and in the determination of the deadline for notifying the Commission: the provider is obliged to notify "without delay and in any event not later than two weeks after that requirement has been met or is known to be met". The dies a quo, therefore, can be the time when a certain requirement is fulfilled, or the time when it is known (how?) that it will be fulfilled. The time limit admits three alternatives: (1) without delay (what does it mean?), (2) two weeks from when it is known that a certain requirement will be fulfilled, or (3) two weeks from when a certain requirement has already been fulfilled. Such indeterminacy is not conducive to the legal certainty that the machinery of Articles 51 and 52 and Annex XIII promote.

What we have called the Commission's review and closure phase involves several ways in which the Commission can determine that a general-purpose AI model presents systemic risks: (1) if it considers that the provider has not been able to demonstrate the absence of risks once the objective parameters requiring it to notify have been met (Art. 52.3 AIA); (2) if it designates ex officio when it is aware that it presents systemic risks and has not notified it (Art. 52.1 and Recital 113 AIA); or (3) if it designates ex officio following a qualified alert by a group of independent scientific experts, when the requirements of Annex XIII are met (Art. 52.4 and 90.1.b AIA).

The legal certainty of this designation procedure ends with a consideration, perhaps not entirely explicit in the articles: an AI model in general use must be considered as systemically risky as soon as (1) the Commission receives the notification from the provider, or (2) it designates the model as a systemically risky model in accordance with the three avenues described above. The rule does not clarify the value of silence. It can be assumed that the provider is to be considered as a provider of a general-purpose AI model with systemic risk: (1) as soon as it notifies the Commission that it has exceeded the objective thresholds without providing arguments challenging the rating; (2) as soon as it notifies the Commission, providing arguments challenging the rating, and the Commission does not reply, or replies in the negative; and (3) as soon as the Commission notifies it of its designation. The classification of a general-purpose AI model as a model with systemic risk entails a series of obligations detailed in art. 55 AIA. This means that the provider must have a minimum of legal certainty, which would be lessened if these rules could not be enforced without some kind of communication with the provider, like the ones listed above. These obligations do not include, as is not generally the case, the obligation for the model to produce reliable results[33].

It is interesting to note from the AIA definition not only what general-purpose AI models are, but also what they are not: AI models "used for research, development or prototyping activities prior to commercialisation" are not considered to be general-purpose AI models (art. 3.63 AIA *in fine*). This is intended to ensure that the AIA does not undermine research and development activities, in line with its declaration of support for innovation and respect for the freedom of science (Recital 25). This implies that AI systems and models *specifically* developed and put into service *solely* for scientific research and development purposes are excluded from the scope of AIA

[33] Watcher et al., op. cit., p. 40.

(Recital 25 AIA)[34]. It must be distinguished whether scientific research and development purposes are the only possible use of the AI system or model or whether it can be used, inter alia, for scientific research and development activities. In the latter case the AIA will not apply before its introduction on the market, but once it is introduced on the market the content of the AIA will be fully applicable to it (Art. 2.8 AIA).

Market introduction is defined as "the first placing on the Union market of an AI system or a general-purpose AI model" (Art. 3.9 AIA[35]). Also, where the provider of a general-purpose AI model integrates a proprietary model into a proprietary AI system that is placed on the market or put into service, it is to be considered as having been placed on the market (Recital 97). Tests conducted under real conditions that meet the requirements of Articles 57 or 60 (Art. 3.57 AIA)[36] shall not be considered as placed on the market. general-purpose AI models already operating on the EU market during the first year after the entry into force of the AIA will have 36 months to comply with its requirements (Art. 111.3 AIA).

IV. Some regulatory challenges of generative Artificial Intelligence

The use of generative AI poses undoubted advances, but it also presents risks in a number of areas such as civil liability arising from the use of AI, the right to receive accurate information, intellectual property, data protection.

Civil liability arising from the use of AI has to be seen in the light of two Directives that are still at the proposal stage: the new Product Liability Directive, and the AI Liability Directive. To the extent that liability aspects arising from the use of AI fall outside the AIA we will not deal with this aspect[37].

As regards the right to receive truthful information, the AIA highlights that AI models in general use can pose systemic risks, such as disinformation (Recital 110 AIA), which can jeopardise democratic and electoral processes

[34] Emphasis added.

[35] Cf. also Art. 3.2 EU Regulation 2019/1020.

[36] Articles 57 and 60 are part of Chapter VI (Articles 57 to 63), which deals with innovation support measures. Article 57 sets out requirements for controlled AI test sites, and Article 60 sets out requirements for testing of high-risk AI systems under real conditions outside controlled sites.

[37] It has been dealt with in Spain in general terms by Muñoz García, C. (Regulación de la inteligencia artificial en Europa: incidencia en los regímenes jurídicos de protección de datos y de responsabilidad por productos, Tirant lo Blanch, Valencia, 2023), or Navas Navarro, S. (Daños ocasionados por sistemas de inteligencia artificial: especial atención a su futura regulación, Comares, Granada, 2022), and specifically regarding Generative AI Novelli et al., op. cit.

(Recital 120 AIA), as well as large-scale manipulation, fraud, impersonation, and consumer deception (Recital 133 AIA). The risk of disinformation is increased by ultra-fakes: AI-generated or manipulated images, audio or video that resemble real persons, objects, places or other entities or events and are likely to mislead a person into believing them to be genuine or truthful (Art. 3.60 AIA)[38]. The EU has been developing strategies against disinformation since at least 2017[39], culminating to date in a strengthened Code of Best Practice on disinformation in 2022[40]. The treatment of this issue goes beyond the AIA and the AI models in general use, so, as in the previous case, we simply point it out.

In terms of intellectual property challenges, we can highlight as challenges not only the possible impediments to reuse of material by general-purpose AI models (to which the AIA pays some attention, as we will see below), but also the necessary distinction between a use of the AI model as a mere tool, and a creative use of the model[41]. As far as the reuse of prior art is concerned, apart from the provisions on the subject in the AIA, it is not to be expected that these cases will be prosecuted, at least as far as the generative AI tools of large companies are concerned, which have already announced financial compensation for possible infringements that may have been committed.[42]

Regarding the first question, both the AIA and Directive 2019/790 offer the following answers: results produced by general-purpose AI models are subject to a similar regulatory logic as derivative works (Recital 105 AIA), which leads to a number of obligations for providers of general-purpose AI models in coordination with Art. 4.3 of EU Directive 2019/790. Art. 4 of EU Directive 2019/790 mandates Member States to provide for exceptions to certain copyright[43] in relation to reproductions and extractions of works and other subject matter that are legitimately accessible for text and data mining purposes, provided that the use of the works and other subject

[38] The definition was introduced by the European Parliament's amendment 203 to the original AIA Proposal, although the term is already mentioned since 2021.

[39] Commission, Communication "Fighting online disinformation: a European approach", Brussels, 26 April 2018, COM (2018) 236 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236.

[40] https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation

[41] Novelli et al., op. cit., p. 14.

[42] CMA - Competition & Markets Authority (2024): AI Foundation Models - Technical update report, https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf, p. 54.

[43] Arts. 5.a and 7.1 of Directive 96/9/EC; 2 of Directive 2001/29/EC; 4.1.a and b of Directive 2009/24/EC; and 15 of Directive 790/2019/EU.

matter is not expressly reserved by rightholders in an appropriate manner in accordance with Art. 4.1 and 3 of EU Directive 2019/790. In the European Parliament's Amendment 399, the proposed Art. 28.b envisaged three obligations for providers of foundational generative AI models: (1) transparency obligations (informing persons that they are interacting with an AI system)[44]; (2) design and develop the model in a way that ensures safeguards against the generation of content that infringes intellectual property law; and (3) document and make publicly available a detailed summary of the training data. The latter obligation is generally applicable for general-purpose AI models under Art. 54 AIA.

The second question (as to when a result produced with generative AI should be considered eligible for protection) does not allow for an *a priori* solution but will need a case-by-case examination. However, criteria can be offered that are not new but are connected with the classic distinction between photography and mere photography of art. 128 TRLPI[45]: we will speak of intellectual (or industrial) property works whenever the AI tool is used as a mere instrument, so that it is possible to recognise a genuine human activity in the choices of both results and instructions introduced in the AI tool; a question that will not arise if the general-purpose AI model works autonomously[46]. It remains to be determined, in each case, what constitutes genuine and recognisable human activity as such.

As regards the data processing aspects, it should first of all be recalled that generative AI is a tool fed with large amounts of data, both personal and non-personal. It should also be recalled that the notion of personal data and the processing of personal data is expansive[47]. This necessarily implies that there is a risk of infringement of data protection law along the entire data value chain (from the collection and processing of data to even the results obtained from their processing) and reflects the importance of data protection by design and by default as provided for in Art. 25 of the GDPR. Even personal information can be inferred from reverse engineering, which leads

[44] Watcher et al., op. cit., p. 40.

[45] On this subject, Bondía Román, F., "Los derechos sobre las fotografías y sus limitaciones", Anuario de Derecho civil, Tomo LIX, Fasc. III, July-September 2006, https://www.boe.es/biblioteca_juridica/anuarios_derecho/abrir_pdf.php?id=ANU-C-2006-30106501114, pp. 1065-1114.

[46] Novelli et al., op. cit., pp. 18-19.

[47] Romeo Casabona, C., "Datos personales (Comentario al artículo 4.1 RGPD)", in Troncoso Reigada, A. (dir.), Comentario al Reglamento General de Protección de Datos y a la Ley Orgánica de Protección de Datos Personales y Garantía de los Derechos Digitales, Aranzadi, Navarra, 2021, pp. 574.

to the need to consider differential privacy techniques[48] that might even fall short in this technological context.

Data processing issues concern (1) whether and (2) how personal data should be processed in the training of generative AI tools. Seven issues related to data processing can be highlighted[49]: (1) the legitimacy basis for processing data in model training; (2) the legitimacy basis for processing data in the case of prompts[50]; (3) information requirements; (4) issues related to model inversion, data leakage and exercise of the right to erasure; (5) automated decisions[51]; (6) protection of minors; and (7) respect for the principles of purpose limitation and data minimisation. We will focus on the first of these[52].

As regards the legitimisation basis for the processing of personal data in training, it should be recalled that all data processing must be carried out in accordance with at least one of the legitimisation bases set out in Art. 6 GDPR, even in the case of companies established outside the EU but offering services in the EU[53] and prior to the market introduction of the model. While consent is at the outset the most legally certain basis of legitimation for the controller, this is not the case for Generative AI because of the enormous costs for the controller to ensure that all data subjects have given specific, free, unambiguous and informed consent for the huge variety of data processing activities that will take place[54]. It therefore seems more appropriate to rely on legitimate interest as a basis for legitimisation, provided that the bal-

[48] Differential privacy is a mathematical method that allows for better privacy protection by incorporating sufficient random noise into the original information. The result does not lose value by application of the law of large numbers, but the introduction of noise allows for a plausible deniability that a particular person's data is part of the analysis set (Dwork, C., "Differential privacy", in Bugliesi, M. et al. (eds.), Automata, Languages and Programming, Springer, Berlin-Heidelberg, 2006, pp. 1-12).

[49] Novelli, et al., op. cit., pp, 7-14.

[50] In many cases, the use of generative AI tools involves a kind of dialogue with the tool, in which you can "tell" it information about another person without their consent - information that it can use for its own training.

[51] Following the broad conception of the notion of "decision" defended by the CJEU in the SCHUFA case, one might wonder whether we are not in this case also dealing with fully automated decisions under Art. 22 GDPR (Novelli et al., op. cit., p. 12).

[52] For a detailed review, Novelli, C. et al., "Generative AI in EU Lawcit…" pp, 7-14. On the other hand, in this work, Jiménez López, J., "Protección de datos y Reglamento de Inteligencia Artificial.

[53] Art. 3.2.a RGPD.

[54] There have even been situations where tools such as Chat GPT 3.5 have provided explicit consent lists for the use of data incorrectly (Watcher, S.; Mittlestadt, B.; Russell, C., "Do large language models have a legal duty to tell the truth?", Royal Society Open Science, May 2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4771884, p. 11, note 88).

ancing test between the legitimate interests of the data controller (the AI tool manager) and the fundamental rights and freedoms of data subjects is passed; an analysis that should take place on a case-by-case basis, i.e., tool-by-tool[55]. The potential usefulness of using synthetic data has also been highlighted, although for the time being the capacity to use synthetic data on a large scale does not allow it[56], in addition to the fact that, according to Art. 4.1 GDPR, the synthetic origin of the data would not prevent them from being classified as personal data.

The issue becomes somewhat more complicated if we take into account that not only the concept of personal data is expansive, but also the concept of special categories of personal data in Art. 9.1 GDPR. This could be observed in the CJEU of 7 July 2023, which understands that even those data that allow the disclosure of information falling under one of the special categories of data of Art. 9.1 GDPR are already special category data regardless of whether the information disclosed is accurate or not[57]. In Art. 9.2 GDPR, which contains the exceptions to the general prohibition of processing of special categories of data, there is no equivalent to legitimate interest. Exceptions could be explored such as the one concerning research activities (Art. 9.2.j GDPR) or the one concerning data that the data subject has manifestly made public (Art. 9.2.e GDPR)[58]. As regards the latter, it must be taken into account whether the data subject explicitly intended, through a clear affirmative action, to make these data public[59], and even so, the exceptions of Art. 9.2 GDPR must be interpreted restrictively. Thus: (1) the data must not relate to persons other than the one who made them public[60]; (2) the mere consultation of websites cannot be understood as data that the data subject manifestly makes public[61]; and (3) the reasonable expectations of the data subject (i.e., whether he could expect that his data would be used to train AI

[55] Gil González, E.; De HerT, P., "Understanding the Legal Provisions That Allow Processing and Profiling of Personal Data-an Analysis of GDPR Provisions and Principles", ERA Forum 2019, vol. 4, 2019, https://research.tilburguniversity.edu/en/publications/understanding-the-legal-provisions-that-allow-processing-and-prof, pp. 618-619; Novelli et al., op. cit., p, 8.

[56] CMA, op. cit., p. 41.

[57] CJEU of 7 July 2023 (C-251/22), cons. 68-73.

[58] Vid. ECDC (2024), Report of the work undertaken by the ChatGPT Taskforce, https://www.edpb.europa.eu/our-work-tools/our-documents/other/report-work-undertaken-chatgpt-taskforce_en, p. 7, cons. 18.

[59] CJEU of 7 July 2023 (C-251/22), cons. 77.

[60] CJEU of 7 July 2023 (C-251/22), cons. 75.

[61] CJEU of 7 July 2023 (C-251/22), cons. 79.

models and tools) at the time of collection of his data[62] must be taken into account. As regards research activities, it is worth noting the AIA's stated support for innovation, respect for the freedom of science and its willingness not to undermine research and development activity (Recital 25 AIA)[63]. However, it is still the GDPR that delimits what is 'research' with regard to data processing. Although it favours a broad concept of research[64], it excludes activities intended for commercial exploitation (Recital 159 and 162 GDPR). All this allows affirming the convenience of designing a new exception in the framework of Art. 9.2 GDPR referring to the use of personal data for the training of AI models and systems of general use with adequate safeguards to preserve the balance between the social interest in the benefits derived from AI training and the protection of the rights and freedoms of citizens[65].

V. Applicability of the Regulation as a general rule and regulatory developments in the treatment of general-purpose Artificial Intelligence

The harmonised rules of the AIA in respect of high-risk AI systems are general rules and, therefore, should be understood without prejudice to those related to data protection, consumer protection, fundamental rights, employment, worker protection and product safety (Recital 9 AIA). In particular, with regard to general-purpose AI models, the AIA does not affect EU copyright law (Recital 108).

The AIA is also a rule for the introduction and monitoring of products on the EU market, as we can see in its Art. 1 and 2 and in Recital 118. Art. 1.2.e AIA announces precisely that harmonised rules are established for the introduction of AI models for general use on the EU market, rules that are applicable to providers intending to introduce them on the EU market regardless of whether they are based in the EU or in a third State (Art. 2.1.a AIA). Therefore, it must be interpreted in accordance with the rules that refer to this aspect such as EU Regulations 765/2008 and 1020/2019 or Decision 768/2008/EC (Recital 9 AIA), and especially with EU Regulation 2019/1020

[62] CJEU of 7 July 2023 (C-251/22), para. 117.

[63] The wording of this recital is almost identical to that of the European Parliament's Amendment 11.

[64] Martín Urganga, A., "Protección de datos y fomento de la investigación científica: la necesidad de un equilibrio adecuado", in TRONCOSO REIGADA, A. (dir.), Comentario al Reglamento General de Protección de Datos y a la Ley Orgánica de Protección de Datos Personales y Garantía de los Derechos Digitales, Aranzadi, Navarra, 2021, p. 1221.

[65] Novelli et al., op. cit., p. 9.

on market surveillance and product conformity. So much so that Art. 18 of EU Regulation 2019/1020 acts as a general rule on the procedural rights of providers (Recital 164 *in fine* AIA).

It is important to note in this respect that many AI systems and models are introduced to the EU market directly in the digital environment, e.g., on very large platforms or search engines (VLOP/VLOSE). In these cases the AIA complements the provisions of the Digital Services Act (DSA) on risk management. In particular, very large search engines and platforms must carry out an assessment of systemic risks arising from the design, operation, and use of their services and, where necessary, take appropriate mitigating measures. The requirements of the AIA applicable to AI systems and models are presumed to be fulfilled if the DSA has already been complied with unless systemic risks other than those covered by the DSA are identified (Recital 118). Art. 34.1 DSA details a list of potential systemic risks of VLOP/VLOSE, some of which may be (partially) reflected in Annex III of the AIA. However, this does not seem to be a closed list if we take into account not only the open way in which systemic risks are described in Art. 34.1 DSA, but also the foreseen publication of annual reports by the Digital Services Board in cooperation with the Commission, including the detection and assessment of the most prominent and recurrent systematic risks reported by VLOPs/VLOSE (Art. 35.2 DSA). Therefore, it seems that Recital 118 DSA is thinking of systemic risks that are circumstantially detected after the information has been submitted in compliance with the DSA and specifically referred to in the AIA.

As mentioned above, the rules on general-purpose AI models were introduced after the bulk of the AIA proposal had been drafted and after the emergence of tools such as ChatGPT. As a result of this last-minute introduction of the rules on general-purpose AI is also the distribution of supervisory competences (Recital 161): if an AI system is based on a general-purpose AI model and both are from the same provider, supervision will be carried out at EU level by the AI Office[66], which will have market surveillance authority powers in accordance with EU Regulation 2019/1020[67]. In all other cases, national market surveillance authorities will be in charge of supervision, but where a general-purpose AI system can be used directly by those responsible for its deployment for purposes considered high risk, national au-

[66] The AI Office is established by Commission Decision (DOIA) of 24 January 2024 (Art. 3.47 and Art. 64 AIA). Decision available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32024D01459

[67] Vid. art. 3.4 and 10 EU Regulation 2019/1020.

thorities should cooperate with the AI Office (AIO)[68] under the cross-border mutual assistance procedure provided for in Chapter VI of EU Regulation 2019/1020 (Articles 22 to 24).

Before closing this section on the evolution of the regulation of general-purpose AI systems and models, a brief reference should be made to the European Parliament's amendment 213, which in the end did not make it into the final version of the AIA. This amendment proposes the introduction of an article on general principles applicable to all AI systems, a sort of homologue (with a few exceptions) to Article 5 of the GDPR. As general principles, they would have been applicable to all AI systems and models, including those in general use. According to this proposal, operators of AI models and systems should, in the interest of promoting a coherent, human-centred, European approach to ethical and trustworthy AI, comply with the following principles: (1) human intervention and vigilance; (2) technical soundness and security; (3) privacy and data governance; (4) transparency (one should add, explainability); (5) diversity, non-discrimination and fairness; and (6) social and environmental welfare. This is not the place to reflect on whether it should have been these six principles or others[69], but at least to regret that such an article has not been introduced in the end, given the vast amount of thoughtful development on ethical principles of AI.

VI. Obligations of general-purpose Artificial Intelligence providers in the Regulation

We can group the obligations of general-purpose AI providers into three: (1) pre-marketing obligations (art. 54 AIA); (2) general (art. 53 AIA); and specific to providers of general-purpose AI models involving systemic risk (art. 55 AIA). In addition, it should be borne in mind that the Commission may request from the provider of the general-purpose AI model not only the documentation referred to in articles 53 and 55 AIA, but also any other documentation it considers necessary to assess the provider's compliance with the AIA (art. 91.1 AIA).

The original proposal for obligations for providers of general-purpose

[68] The 2021 AIA Proposal refers to the European AI Committee. This name changes following the European Parliament's Amendments 524 et seq.

[69] For this, see Cotino Hueso, L., "Ética en el diseño para el desarrollo de una inteligencia artificial, robotica y big data confiables y su utilidad desde el derecho", Revista catalana de dret públic, n.º 58, 2019, https://revistes.eapc.gencat.cat/index.php/rcdp/article/view/10.2436-rcdp.i58.2019.3303/n58-cotino-es.pdf, pp. 36-40.

AI models (then called "foundational models") is contained in the European Parliament's Amendment 399, which proposed a new Article 28.b.2, containing pre-marketing and post-marketing obligations for the model. The system has changed considerably in the final text: the pre-marketing obligations are only applicable to third country providers (Art. 54.1 AIA), the need to appoint an authorised representative in the EU and to cooperate with AIO have been added. Regarding post-marketing obligations, the obligation to keep the documentation up to date for ten years (in relation to the deadline, Art. 18.1 AIA) is maintained, and the obligation to prepare and publish a detailed summary of the training content of the tool is added.

Firstly, they will have to appoint authorised representatives established in the EU (Art. 54 AIA), with a minimum content of the mandate: (1) to check that the technical documentation has been drawn up, a copy of which must be kept at the disposal of AIO and the competent national authorities; (2) that the provider complies with the obligations of Art. 53 and, where applicable, 55; (3) to provide AIO, upon reasoned request, with information demonstrating compliance with these obligations; (4) and to cooperate with AIO and national authorities in the preparation of the technical documentation. 53 and, where applicable, 55; (3) provide AIO, upon reasoned request, with information demonstrating compliance with these obligations; (4) and cooperate with AIO and national authorities in actions undertaken with general-purpose AI models with systemic risk.

Article 53 AIA contains a list of general obligations, which can be grouped into three types: (1) documentary (2) relating to intellectual property, and (3) cooperation. It does not include an obligation to ensure that AI models in general use provide reliable results[70].

With regard to documentary obligations, providers of general-purpose AI models must: (1) develop and maintain up-to-date technical documentation of the model (including information on the training process, testing and evaluation results); (2) develop and maintain up-to-date information and documentation to enable providers of AI systems intending to integrate the general-purpose AI model to understand the capabilities and limitations of such model in order to comply with the AIA; and (3) develop and make publicly available a sufficiently detailed summary of the content used for training the model, in accordance with the model provided by the AIO (arts. 53.1. a, b and d, and 53.7 AIA).

Open source general-purpose AI models are not required to comply with the documentary obligations of Art. 53.1(a) and (b) AIA, but must make

[70] Watcher et al., op. cit., p. 40.

available to the public a detailed summary of the content used for their training in accordance with Art. 53.1.d AIA. The documentary obligations of Art. 53.1 (a) and (b) AIA must be complied with as specified in Annexes XI and XII AIA respectively. These Annexes, as elsewhere in the AIA, may be updated in line with technological developments by means of delegated acts under Art. 97 AIA (Art. 53.5 and 6 AIA respectively). Finally, all information and documentation prepared under Art. 53 is subject to a duty of confidentiality under Art. 78 AIA (Art. 53.7 AIA). To the extent that such documentation includes personal data, the duty of confidentiality under Art. 78 AIA must be supplemented by Art. 5.1.f GDPR.

With regard to documentary obligations, this should include those relating to the development of documentation (which should be kept up to date) on the general-purpose AI model by downstream providers, as part of the responsibilities of providers of general-purpose AI models along their value chain. It should be borne in mind that general-purpose AI models may form the basis of downstream systems supplied by other providers who will therefore need to have a good understanding of the models and their capabilities, both for technical reasons and in order to comply with the AIA and other regulations (AIA Recital 101).

As far as copyright obligations are concerned, these are not particularly clearly set out in Art. 53.1.c AIA, which we have already discussed above[71].

With regard to the duties of cooperation, providers of general-purpose AI models must cooperate with the Commission and the competent national authorities in order to facilitate compliance with the AIA in accordance with Article 53.3 of the AIA. This generic duty of cooperation is specified in various parts of the AIA, and is not limited to the relationship between providers of general-purpose AI models and authorities, but also between authorities and each other, in the framework of the principle of loyal cooperation in Article 4.3 TEU. Thus, the duties of cooperation with the AIO in the framework of Art. 75.2 AIA, as well as the powers of review and supervision of both the AIO (Art. 75.1 and 3 AIA) and the Commission (Art. 88.2 AIA) must be taken into account. Also, as regards general-purpose AI systems, to the extent that they may be used as high risk AI systems either on their own or as parts thereof, providers of general-purpose AI systems must cooperate closely with the relevant providers of high risk AI systems (Art. 85 AIA).

In addition to the general obligations established for providers of general-purpose AI models, when these are qualified as general-purpose AI models

[71] See above, supra, para. on normative challenges of generative AI.

with systemic risk, they must comply with the specific obligations established in Art. 55 AIA.

The first of these is to assess models with a view to detecting and reducing systemic risk (Art. 55.1.a AIA). This assessment should be carried out in accordance with standardised protocols and tools according to the state of the art and include adversarial simulation testing with the model. Adversarial simulation tests or robustness tests allow identifying vulnerabilities through the simulation of attacking systems within a network and suggesting improvements that allow a better understanding and continuous improvement of the model and reinforce its security and reliability[72]. They should also identify the source, assess and mitigate systemic risks at EU level that may arise from the development, market introduction, or use of the general-purpose AI model with systemic risk (Art. 55.1.b AIA); monitor, document and report without undue delay to the AIO information regarding serious incidents and possible remedial actions; and also ensure that an adequate level of protection of the model's cybersecurity and physical infrastructure is in place (Art. 55.1.c and d AIA).

The duty of confidentiality in Art. 53.7 AIA is reiterated in Art. 55.3 AIA, with the same wording. The question arises as to the need for Art. 55.3 AIA if Art. 53 is already generally applicable to all providers of general-purpose AI models, whether or not they are systemically risky models. At last, codes of practice allow providers of general-purpose AI models with systemic risk to demonstrate compliance with their obligations (Art. 55.2 AIA).

Finally, codes of good practice serve as a crucial instrument to ensure that providers of general-purpose AI models properly fulfil the obligations arising from the AIA, while also simplifying the process of demonstrating compliance with these obligations (Recital 117 and Arts. 53.4 and 55.2). This is an introduction that was not foreseen in the 2021 Proposal nor in the European Parliament's Amendments. On the other hand, for the purposes of the AIA, harmonised standards refer to all technical specifications that have been adopted by a recognised standardisation body. While compliance with these standards is not mandatory, they are of repeated or continuous application. These standards were developed in response to a request from the Commission, as stated in art. 3.27 of the AIA and 2.1.c of EU Regulation 1025/2012. It is always up to the provider to use alternative methods other than codes of practice and harmonised standards.

[72] Hannon, B.; Kumar, Y.; LI, J. J.; Morreale, P., "From Vulnerabilities to Improvements-A Deep Dive into Adversarial Testing of AI Models", Congress in Computer Science, Computer Engineering & Applied Computing, 2023, pp. 2645-2649.

VII. Supervision and monitoring, and sanctioning regime

The Commission is entrusted with powers to supervise and monitor compliance with the obligations of providers of general-purpose AI models, the execution of which it delegates to the AIO, which should be able to take the necessary measures to supervise the effective implementation and enforcement of the obligations of providers of general-purpose AI models laid down in the AIA, and to this end may request information, evaluate and impose measures on providers of general-purpose AI. It may also rely on advice from the group of scientific experts provided for in the AIA, who must be selected on the basis of scientific or technical expertise in the field of AI and carry out their functions impartially and objectively (Recital 162 and 164, and Arts. 68 and 88 AIA).

The AIO has supervisory powers when an AI system is based on a general-purpose AI model and the same provider develops both model and system, being considered for these purposes as a market authority in accordance with EU Regulation 2019/1020 (art. 75.1 AIA). It is also competent to take measures in relation to the implementation of and compliance with the AIA by providers of general-purpose AI models, as well as for the observance of approved codes of practice (Art. 89.1 AIA); and to carry out assessment tasks in the framework of Art. 92 AIA.

With regard to the sanctioning regime in the AIA, it can be divided into two categories: adoption of measures and sanctions as such, the former in the hands of the Commission and the latter in the hands of the Member States and the EDPS (in the case of conduct contrary to the GDPR). The actual determination of sanctions remains in the hands of the Member States, except for fines for providers of general-purpose AI models acting intentionally or negligently, which according to Art. 101 AIA are imposed by the Commission (Art. 101.1 AIA).

The Commission is competent to: (1) request the adoption of measures to ensure timely compliance with the obligations of providers of general-purpose AI models; (2) require a provider to implement risk mitigation measures when the assessment under Art. 92 AIA indicates the existence of an EU-wide systemic risk; and (3) restrict the marketing of the model (Art. 93.1.a, b and c AIA). Before requesting action, AIO can enter into a structured dialogue with the provider of the general-purpose AI model, aimed at avoiding unilateral action by the Commission, because if during the structured dialogue the provider commits to take systemic risk mitigation measures, the Commission can adopt a decision making the provider's commitments binding and declaring that there are no grounds for action (Art. 93.2 and 3 AIA).

With regard to fines, they must be of an "appropriate level" (Art. 169 AIA), and be "effective, proportionate, and dissuasive" (Art. 101.3 AIA) taking into account: (1) the nature, gravity, and duration of the infringement (2) the principles of proportionality and appropriateness, and (3) the commitments made by providers under Art. 93.3 AIA or adherence to codes of good practice. The AIA sets a maximum amount of fines that may be imposed by the Commission. The maximum amount of these fines is 3% of the total worldwide turnover for the preceding financial year or 15 million euros. The cases in which the Commission may impose fines are: (1) breach of the rules of the AIA; (2) failure to comply with a request for documentation or information under Art. 91 AIA or the provision of inaccurate, incomplete, or misleading information; (3) failure to comply with a measure requested under Art. 93; or (4) failure to provide the Commission with access to the generally available Model AI for the purpose of an assessment under Art. 92 AIA.

The imposition of fines must be carried out in accordance with certain procedural rules set out in Art. 101. Both the Commission's and the CJEU's actions, if we follow the literal wording of the AIA, will be of their own initiative when the circumstances foreseen are met. However, we could ask ourselves whether it is possible for individuals to denounce infringements of the AIA or of the Commission in the imposition of fines, even though this is not expressly provided for in the rule. In other words, is it sufficient for an EU rule to be invoked if it sets out a clear and unconditional obligation, or must the possibility of individual whistleblowing also be provided for? The answer to this question was answered by the well-known judgment of the CJEU of 17 September 2002, which held that the guarantee that EU law will be enforceable also requires that it can also be enforced in civil proceedings brought by private individuals[73]. This is consistent with the assumption that it is the national judge who is responsible for the enforcement of EU law in each Member State. This criterion has been proposed with respect to recent EU rules that also do not explicitly provide for the possibility for private citizens to bring actions to enforce them, as is the case with EU Regulation 2019/1150[74]. In our opinion, it seems reasonable to understand that this criterion is equally applicable to the AIA, which may be particularly interesting during its first years of operation, when it is foreseeable that the different implementing rules foreseen are still being drafted.

[73] CJEU of 17 September 2002 (C-251/22), cons. 30 y 31.

[74] Jens-Uwe, F., "Individual Private Rights of Action under the Platform-toBusiness Regulation", European Business Law Review, vol. 34, volume 4, 2023, pp. 559-560.

VIII. Conclusions

The emergence of tools such as Chat GPT at the end of 2022 has caused a social and also a regulatory earthquake. A good example of the regulatory earthquake is the introduction of specific provisions relating to general-purpose AI models in the final version of the AIA, incorporating (and extending) the amendments tabled by the European Parliament in June 2023 to the 2021 AIA Proposal, which for the first time introduce specific provisions relating to the then so-called "foundational models". The huge impact of these tools is also illustrated by the fact that the need and manner of their inclusion in the text of the AIA was one of the last critical points of discussion in the negotiation that took place at the end of 2023.

It is said that general-purpose AI models are AI models that have been trained with a lot of data through large-scale monitoring and are pretty general. They can do a lot of different tasks well and can be added to a lot of different systems or applications (Art. 3.63 AIA). An AI system is considered to be general-purpose when it is based on a general-purpose AI model (Art. 3.66 AIA). Within general-purpose AI models, two subcategories should be highlighted: generative AI models, which are those capable of flexibly generating text, audio, image, or video content; and general-purpose models that present systemic risk.

A general-purpose AI model presents systemic risk if it has high impact capabilities (according to Annex XIII AIA, which may be updated according to technical developments), or if it has a significant potential impact on the internal market due to its scope. Providers of AI models presenting systemic risk will be subject to additional obligations, as outlined in Articles 53 and 54 AIA, which pertain to general and specific obligations for AI models with systemic risk. This justifies a certain need for legal certainty for these providers, which should be considered in this category, provided that they exceed certain standards and so inform the Commission or when the Commission informs them that they have this status because they have exceeded certain standards.

Generative AI models, on the other hand, present certain challenges related to civil liability arising from the use of AI, misinformation[75], personal data protection and intellectual property[76]. As regards the processing of per-

[75] Watcher et al., op. cit., p. 5.

[76] BSI - Federal Office for Information Security of the German Government (2024): Generative AI Models - Opportunities and Risks for Industry and Authorities, https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Generative_AI_Models.html, p. 9.

sonal data, the issues relate to whether and how personal data should be treated both in the training of AI tools and subsequently in their use. It should be recalled that general-purpose AI models are characterised by the use of large amounts of data, which raises questions about the basis of legitimacy (and, where appropriate, exceptions to the general prohibition of processing special category data) for the training of the model, or concerning the personal data that users themselves enter in the prompts to the tool.

As regards intellectual property aspects, on the one hand, the results produced by general-purpose AI models have a certain similarity with derivative works, although insofar as they are produced through text and data mining, the provisions of Art. 4.3 of Directive 2019/790, to which the AIA refers, must be taken into account. It is also interesting to ask when a result produced by a generative AI tool can be protectable: the criterion to be taken into account, in our view, should still be whether it is possible to recognise genuine human activity in the use of the tool, which translates into asking whether the instructions introduced are sufficiently complex and creative (similar to the classic distinction between a photographic work and a mere photograph). What will be difficult will be to determine in practice when this genuine human activity is recognisable. A final, non-legal but important, question is the social repercussions of the ease of access to tools that make it easier to escape from elaborating complex human thoughts, or at least to escape from the mere discipline of the activity of writing, for example.

The AIA is a standard for the introduction and monitoring of products on the EU market (AI systems and models). This is seen in the cross-references at various points to EU Regulation 2019/1020. This also explains why the obligations laid down for AI models for general use mainly concern information, documentation and cooperation with authorities (Commission, AIO, and national authorities). A distinction is made between three types of obligations for providers of general-purpose AI models: pre-marketing of the model (applicable only to providers not established in the EU), general, and specific to providers of general-purpose AI models presenting systemic risk. These obligations were first introduced by Amendment 399 of the European Parliament; however, they have undergone considerable changes in structure and content. In Amendment 399, the pre-marketing obligations were not restricted to non-EU providers and were somewhat more extensive. However, the general obligations were not as detailed as those in Art. 53 AIA. No specific obligations were foreseen for AI models presenting systemic risks insofar as this category was not included either.

The regulatory system applicable to general-purpose AI models is completed in the AIA with provisions on supervision and monitoring, and penalties. The Commission has been assigned powers of supervision and control of regulatory compliance, the execution of which it delegates to the AIO, which in its operating rules (Decision of 24 January 2024) includes several control provisions in respect of AI models for general use.

The Commission divides the sanctioning regime into two categories: measures and fines. Regarding measures, the Commission possesses the authority to mandate compliance measures, compel a provider to implement risk mitigation measures in cases of systemic risk, and ultimately, restrict the marketing of the model. The AIO plays an important role in this context in that it can engage in structured dialogues aimed at preventing the Commission from acting. Fines must be appropriate, effective, proportionate, and generally dissuasive (Art. 101.3 AIA). In general, it is up to the Member States to set the amount of the fines; but in the case of general-purpose AI models, the Commission will be competent to impose them in accordance with the established ceilings of 3% of the worldwide turnover of the previous financial year or 15 million euros. By way of conclusion, it should be understood that it is possible for individuals to bring a complaint before the ordinary national courts for non-compliance with the AIA, even if this is not expressly provided for in the regulation.

CODES OF CONDUCT, SEALS OR CERTIFICATIONS FOR ARTIFICIAL INTELLIGENCE SYSTEMS THAT ARE NOT HIGH RISK (ARTICLE 95 OF THE AI ACT)

Lorenzo Cotino Hueso

Professor of Constitutional Law at the University of Valencia. Valgrai[1]

I. The regulation in Article 95 of Codes of conduct for non-high-risk systems

The AIA essentially regulates obligations for high-risk systems. However, it also contains some provisions on general-purpose AI models (Chapter V, Articles 51-56) and imposes some transparency obligations on "certain" AI systems in Article 50, Chapter IV.

The European Commission estimates that 90% of AI systems[2] and two thirds of public AI systems will not be classified as high risk.[3] Under the AIA risk model, systems that are not high risk will not be subject to the obligations of the regulation. However, these systems will still be subject to other relevant regulations, such as the GDPR in case of processing of personal data. Furthermore, to ensure product safety, Regulation (EU) 2023/988 of the European Parliament and of the Council of 10 May 2023 on general product safety (Recital 166) will apply as a safety net. This regulation states that it "lays down essential rules on the safety of consumer products placed on the market" (Art. 1.2) and is the regulatory "broom wagon", as it applies to products

[1] cotino@uv.es. OdiseIA. This study is the result of research from the following projects: MICINN Project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by MCIN/AEI/10.13039/501100011033/; "The regulation of digital transformation…" Generalitat Valenciana "Algorithmic law" (Prometeo/2021/009, 2021-24); "Algorithmic Decisions and the Law: Opening the Black Box" (TED2021-131472A-I00) and "Digital transition of public administrations and Artificial Intelligence" (TED2021-132191B-I00) of the Recovery, Transformation and Resilience Plan. CIAEST/2022/1, Research Group in Public Law and ICT, Catholic University of Colombia; CIAEST/2022/1, Digital Rights Agreement-SEDIA Area 5 (2023/C046/00228673) and Area 6 (2023/C046/00229475).

[2] European Commission, Renda. A. (project leader), *Study to Support an Impact Assessment of Regulatory Requirements for Artificial Intelligence in Europe. Final Report (D5),* April 2021. p. 143, https://op.europa.eu/es/publication-detail/-/publication/55538b70-a638-11eb-9585-01aa75ed71a1

[3] JRC, Tangi, L. et al.: *AI Watch European landscape on the use of Artificial Intelligence by the Public Sector*, JRC Science For Policy Report, European Union. 2022, p. 58.

"as long as there are no specific provisions with the same purpose in Union law governing the safety of the products concerned" (Art. 2.1).

For non-high-risk systems, since the initial version of the AIA in 2021[4] a Title IX on "Codes of conduct" has been included, with the objective of voluntary compliance with mandatory requirements for AI systems.[5] These codes could also include voluntary commitments on environmental sustainability, accessibility for people with disabilities, stakeholder involvement in the design and development of AI systems, and diversity in development teams. The intention of the AIA is that non-hazardous systems should be "safe when placed on the market or put into service" (Recital 166). Article 95 has undergone few changes, with the final appearance of the AI Office and the attribution of responsibilities to it, initially assigned to the Commission and the Committee, being relevant. In the final version, additional elements that these Codes may include have been added, linked to the EU Ethical Guidelines, environmental impacts, literacy, inclusive design or harm to vulnerable people.

Given that the implementation of AIA or other obligations would be voluntary and not mandatory, the AIA connects this issue to the field of ethics. For non-high-risk systems, it "may lead to the wider adoption of ethical and trustworthy AI in the Union" (Recital 165). This is the EU's AI brand, also known as ethical AI by design.[6]

Article 95 of the AIA, which was originally a whole chapter and is now Chapter X, "Codes of Conduct and Guidelines", regulates in a rather open-ended normative manner:

- The promotion and facilitation of governance codes and mechanisms by the AI Office and Member States.

[4] The issue was addressed in Recital 81 and Article 69, in addition to the general explanation of the Regulation.

[5] On the subject see: Stuurman, K. and Lachaud, E., *Regulating AI. A Label To Complete the Proposed Act on Artificial Intelligence* January 2022 http://dx.doi.org/10.2139/ssrn.3963890

Galán, C., "The Certification as a Mechanism for Control of Artificial Intelligence in Europe" September 2019. http://dx.doi.org/10.2139/ssrn.3451741 also in "La certificación como mecanismo de control de la inteligencia artificial en Europa" in *bie3: IEEE Bulletin*, no. 14, 2019, pp. 622-637.

Cihon, P. et al. "AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries" in *IEEE Transactions on Technology and Society*, LawAI Working Paper No. 5-2021, https://doi.org/10.1109/TTS.2021.3077595.

[6] An exhaustive analysis can be found in "Ethics in the design for the development of reliable Artificial Intelligence, robotics and big data and their usefulness in law" in *Revista Catalana de Derecho Público* n.º 58 (June 2019). http://revistes.eapc.gencat.cat/index.php/rcdp/issue/view/n58 http://dx.doi.org/10.2436/rcdp.i58.2019.3303

- The voluntary and variable application of obligations from high-risk systems to non-high-risk systems, targeting especially providers, but also deployers.

- In addition to voluntary compliance with the AIA, it is noted that these voluntary Codes may introduce "additional requirements" such as those of the EU Ethical Guidelines, environmental impacts, literacy, inclusive design or harm to vulnerable people.[7]

- Codes should take into account best practices and technical solutions, and should be developed "on the basis of clear objectives and key performance indicators to measure the achievement of those objectives" (Art. 95.2).

- Codes of conduct can be developed by providers, deployers, their representative organisations, civil society and academia,[8] with mention of the interests of SMEs and start-ups.

In addition to the article, other sections of the AIA make some mention of codes and certifications. Thus, it is mentioned that one objective of a sandbox can also be to learn how to apply not only the AIA, but also codes of conduct (Art. 58.2 e). The Board has the task of "issuing written recommendations and opinions" on the development and implementation of codes of conduct and best practices (Art. 66.e)i). Every three years, the Commission should assess the impact and effectiveness of voluntary codes of conduct to promote the application of the requirements for high-risk AI systems to non-high-risk AI systems, and possibly additional requirements (Recital 174). The important Codes of Best Practice for general purpose AI regulated in Article 56 are not addressed here.

II. Finally, the Regulation has not included binding principles for all types of Artificial Intelligence systems.

It should be noted that the EU Parliament's amendments to the AIA in 2023 included the regulation of 'general principles applicable to all AI systems' (Article 4a, new). It was intended to follow the outline of the GDPR when it recognises its essential principles in Article 5. As is well known, for more than thirty years, the "principles" are the fundamental pillars of the data protection legal framework, indeed, they constitute concrete rules applicable

[7] This is stated in Recital 165.

[8] "such as business and civil society organisations, academia, research bodies, trade unions and consumer protection organisations", Recital 165.

to processing operations. So much so that their mere non-compliance directly implies the commission of infringements.

It is worth noting that the EU Parliament's amendments to the AIA in 2023 included the "General principles applicable to all AI systems" (Article 4a, new).[9] This was intended to follow the outline of the GDPR, where the principles of Article 5[10] are fundamental pillars in addition to specific applicable rules whose non-compliance entails infringements.

It was proposed to include principles for all AI systems, high-risk or not, as well as for foundational models. The principles of "human intervention and human oversight" (a),[11] "technical robustness and security" (b),[12] "privacy and data governance" (c),[13] "transparency" (d),[14] "diversity, non-discrimination and equity" (e)[15] and "social and environmental welfare" (f)[16]. It prescribed that "All operators […] shall make every effort to develop and use AI systems or foundational models in accordance with the following general principles which set out a high-level framework to promote a coherent European human-centred approach to ethical and trustworthy Artificial Intelligence".

[9] Amendment 213.

[10] Article 5 GDPR regulates them: lawfulness, fairness and transparency, purpose limitation, adequacy, limitation, necessity and proportionality of data (data minimisation), accuracy, purpose limitation, integrity and confidentiality and proactive liability.

[11] "(a) "Human intervention and monitoring" means that AI systems shall be developed and used as a tool in the service of individuals, respecting human dignity and personal autonomy, and operated in a way that can be adequately controlled and monitored by human beings.

[12] "(b) "Technical robustness and security": AI systems shall be developed and operated in such a way as to minimise unforeseen and unexpected damage, as well as to be robust in the event of unforeseen problems and resistant to attempts to modify the use or performance of the AI system to allow unlawful use by malicious third parties".

[13] "(c) "Privacy and data governance": AI systems shall be developed and used in accordance with existing privacy and data protection standards, and shall process data that meet high standards in terms of quality and integrity".

[14] "(d) "Transparency": AI systems shall be developed and operated by providing adequate traceability and explainability, by making individuals aware that they are communicating or interacting with an AI system, by adequately informing users about the capabilities and limitations of such an AI system and by informing affected individuals of their rights."

[15] "(e) "Diversity, non-discrimination and equity": AI systems shall be developed and used in a way that is inclusive of diverse actors and promotes equal access, gender equality and cultural diversity, while avoiding discriminatory effects and unfair bias prohibited by national or Union law".

[16] "(f) "Social and environmental well-being": AI systems shall be developed and used in a sustainable and environmentally sound manner and for the benefit of all human beings, while monitoring and assessing the long-term effects on people, society and democracy".

It is also true that in that AIA proposal by the EU Parliament, the principles were proclaimed cautiously, if I may say "with the brakes on", modulating and restricting their scope, stating "without creating new obligations under this Regulation".[17] However, it was stated that the principles should inspire standardisation processes and technical guidance.[18]

Finally, the AIA does not include principles, and, for non-high risk systems, Article 95 must essentially be followed. It is important to mention that the Council of Europe AI Convention of 17 May 2024[19] has done so. Chapter III of this Convention deals with "Principles related to activities within the life cycle of Artificial Intelligence systems" and "sets out the common general principles to be applied by each Party […] in an appropriate manner to its domestic legal system".[20] Although with some laxity, the principles regulated in eight articles will have a general projection for all types of AI systems.[21] Obviously, for the States and Parties that sign the Convention and once it enters into force.

[17] Paragraph 2: 'Paragraph 1 shall be without prejudice to the obligations laid down by Union and national law in force. In the case of high-risk AI schemes, the general principles are applied and complied with by providers or implementers through the requirements set out in Articles 8 to 15 of this Regulation, as well as the relevant obligations set out in Chapter 3 of Title III of this Regulation. In the case of foundational models, the general principles are applied and complied with by providers or implementers through the requirements set out in Articles 28 to 28b of this Regulation. For all AI systems, the application of the principles referred to in paragraph 1 may be achieved, as appropriate, through the provisions of Article 28 or Article 52 or through the application of the harmonised standards, technical specifications and codes of conduct referred to in Article 69, without creating new obligations under this Regulation".

[18] 3. The Commission and the AI Office shall incorporate these guiding principles in requests for standardisation as well as in recommendations in the form of technical guidance to assist providers and implementers on how to develop and use AI systems. The European standardisation organisations shall take into account the general principles referred to in paragraph 1 of this Article as performance-based objectives when developing relevant harmonised standards for high-risk AI systems referred to in Article 40(2b).

[19] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, adopted at the 133rd session of the Committee of Ministers, Strasbourg, https://search.coe.int/cm?i=0900001680afb11f. About it, Cotino Hueso, L. "El Convenio sobre inteligencia artificial, derechos humanos, democracia y Estado de Derecho del Consejo de Europa", *Revista Administración & Cidadanía*, EGAP, 2024.

[20] This is stated in the 18 December version as an explanatory note.

[21] It is followed by the version of December 2023, which regulates eight articles (art. 6 to 13) that essentially express and affirm these "principles": human dignity and individual autonomy (art. 6), transparency and oversight (art. 7), accountability and responsibility (art. 8), equality and non-discrimination (art. 9), privacy and protection of personal data (art. 10), preservation of health [and the environment] (art. 11), reliability and trust (art. 12), safe innovation (art. 13).

III. The insertion of Artificial Intelligence in certification models and technology seals in the EU: a Spanish Artificial Intelligence seal?

The AIA incorporates AI into the ecosystem of voluntary certification, seals and codes of conduct promoted by the EU in the technological field. These models involve defining clear standards that organisations must meet in order to obtain certification, verifying compliance through accredited bodies. These instruments make it easier to demonstrate compliance with quality, safety, and ethical standards, increasing consumer and user confidence.

EU regulations support these models to give them credibility and official recognition, especially in sectors such as cybersecurity, regulated by Regulation (EU) 2019/881. Codes of conduct and seals that demonstrate compliance with the GDPR, such as the Luxembourg CNPD's GDPR-CARPA scheme, are also important.[22] In addition, mention can also be made of the electronic trust services certificates issued under the eIDAS Regulation (Regulation (EU) No 910/2014) that guarantee the authenticity and integrity of electronic transactions in the European Union.

Carlos Galán proposed in 2019 the creation of a European Certification Scheme to regulate the development and deployment of AI technologies.[23] It was obviously too early to think about the whole AIA regulatory system.

In Spain, the 2020 National AI Strategy (ENIA)[24] included the development of a code of conduct or "label" as a measure to build trust in AI. This was the first measure within Action Line 6.1, "Building trust in AI", specifically Action 26 "Development of a national AI quality label and the elaboration of a catalogue of supplementary measures to AI Certification at European level". To this end, the Government outlined actions for its implementation through the contract "Sello IA del Gobierno Español"[25]. Devel-

[22] In this regard, CNPD, "The certification scheme GDPR CARPA", at https://cnpd.public.lu/en/professionnels/outils-conformite/certification/gdpr-carpa.html.

CNPD, GDPR-CARPA, *GDPR-Certified assurance report-based processing activities*, Commission nationale pour la protection des données, Luxembourg, 2022, https://cnpd.public.lu/content/dam/cnpd/fr/professionnels/certification/lu-gdpr-carpa-certificationscheme.pdf

[23] Galán, C., "The Certification…" cit. This scheme would have to be backed up by a European regulation that would indicate the standards and technical specifications, the independence of the assessment bodies and that the system would include continuous assessment processes and periodic updates.

[24] SEDIA, *National Artificial Intelligence Strategy*, November 2020, https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/ENIAResumen2B.pdf

[25] Services for the development of Artificial Intelligence impact plans, development of a label and study services related to AI systems experimentation environments. https://planderecuperacion.gob.es/como-acceder-a-los-fondos/convocatorias/PLC/11383932/servi-

opment of the technical standards for the seal or certification; 2. Proposal of the accreditation framework: certification scheme and accreditation process; 3. Guides on best practices for implementing the AI Seal and 4. Development of a software for self-assessment of compliance with requirements

Thus, firstly, the objective is to establish technical requirements aligned with European standards, covering aspects of security and data protection specific to AI. Secondly, for the development of the Spanish Seal, the aim is to develop an accreditation framework in collaboration with entities such as AENOR, UNE or Adigital. In addition, it is considered important to offer alternative certification paths for SMEs and to establish clear procedures for the maintenance and withdrawal of accreditation.

Thirdly, the development of good practice manuals explaining the applicable technical and legislative standards is foreseen. These manuals will address governance, traceability, training, and modelling of algorithms, transparency, explainability, dataset and risk management, and impact assessment. Finally, fourthly, it is planned to develop a software tool for self-assessment of compliance with the Seal's requirements, which will automate and facilitate self-assessment, ensure the persistence and security of information, and provide visual reports on the company's level of AI maturity.

The Spanish plan was under implementation, but the change of government in 2023 seems to have had an impact on this issue. The new National AI Strategy adopted in May 2024[26] almost completely omits references to seals or certificates which are not linked to sustainability.[27]

In terms of experience in the development of certification tools or systems in Spain, *Adigital*'s initiative stands out with its "Certification of Algorithmic Transparency", www.transparenciaalgoritmica.es, launched in January 2024, which evaluates the transparency and explainability of the use of algorithms by companies in Spain. Various concretisations can be accessed on the platform.[28] The systems assessed are high risk and the client must provide evidence (documentation and information) to justify the assessment. Software tools are not used in the process, and the evaluation can be iterative until a

cios-para-el-desarrollo-de-planes-de-impacto-de-la-inteligenciaartificial-desarrollo-de-un-sello-y-servicios-de-estudio-relativos-a-entornos-de-experimentacion-de-sistemas-de-AI

[26] ENIA, May 2024, https://portal.mineco.gob.es/es-es/digitalizacionIA/Documents/Estrategia_AI_2024.pdf

[27] Thus, in "Lever 2: Generating Storage Capacities under Sustainable Conditions", specifically "Initiative 2.3. Without further specification, reference is made to the importance of AESIA for codes in the specific field of Generative AI.

[28] https://www.adigital.org/media/policy-brief_ai-transparency-and-ethics-certifications.pdf

certain score is reached that allows obtaining the transparency certificate. In the summer of 2023, a pilot was conducted with companies such as Adevinta (InfoJobs), Holaluz, and Shakers to test the seal in high-impact environments such as employability and critical infrastructure.[29] The assessment focuses on the product and not on the processes or the management system. In principle, they do not assess legal compliance (such as GDPR). At the end of October 2023, they presented the certificate in Brussels together with the three pilot companies.

The certification model of the *Eticas Foundation*,[30] dedicated to ethics in Artificial Intelligence, assesses the implementation of ethical principles in AI systems.[31] This framework is aimed at companies wishing to certify that their processes and products meet strict ethical criteria, including transparency in algorithms, fairness in outcomes, protection of personal data, accountability in automated decisions, and explainability of AI processes. It has an initial focus on Europe, but it is open to organisations worldwide. Although no concrete information on the system is available, the certification includes elements on transparency in algorithms, fairness in outcomes, protection of personal data, accountability in automated decision-making, and explainability of AI processes. In addition, it requires the implementation of mechanisms for continuous monitoring and evaluation of the ethical impact of the systems.

IV. The various European and international Artificial Intelligence initiatives and certification tools or labels

The possibilities for AI certification in the coming years are numerous and diverse. Government-driven certification models, public-private partnerships and private models may coexist in the market. These systems will focus on specific sectors such as business, public, education, health, environmental, inclusion, media, and generative AI. Products will also be developed that focus on traceability compliance, transparency, oversight, quality, and data governance, with seals of varying levels of stringency.

A variety of certification solutions and tools already exist. The OECD has a comprehensive repertoire which, as of May 2024,[32] includes more than

[29] https://www.adigital.org/actualidad/adigital-arranca-su-certificacion-de-transparencia-algoritmica-con-las-tres-primeras-empresas-acreditadas/

[30] https://eticas.ai/guide-to-algorithmic-auditing

[31] https://eticas.org/

[32] https://oecd.ai/en/catalogue/tools?approachIds=3&approachIds=2&approachIds=1&toolTypeIds=20&toolTypeIds=21&orderBy=dateDesc&toXLS=null&page=1

30 AI certification models.[33] Only a few that have stood out at the time are highlighted here, although it is not easy to know how up to date they are.

Thus, in addition to the ISO or NIST initiatives, which are of the utmost relevance, the following are now mentioned: *ALTAI* model*, Future-AI* initiative*, capAI, AI Safety Institute, Ada Lovelace Institute, RIAL* initiative*, Fairly Trained* Certificate*, German AI Association, AI Cloud Service Compliance Criteria Catalogue (AIC4), IEEE CertifAIEd, KI Bundesverband AI* Seal of *Approval, Towards Auditable AI Systems* certification model*, Denmark's new labelling program for IT security and responsible use of data* or the *Responsible Artificial Intelligence Institute (AIA Institute).*

Of particular note is the *ALTAI* model, known from the EU Expert Group's Ethical Guidelines for Trusted AI, with its comprehensive checklist for evaluating a design ethics model[34]. This model, developed by the vice chair of the AI HLEG and his team at the *Insight Center for Data Analytics* at University College Cork[35] guides AI developers and deployers through an accessible and dynamic checklist, focusing on seven key requirements: human agency and oversight, technical robustness and security, privacy and data governance, transparency, diversity, non-discrimination and equity, environmental and social well-being, and accountability.[36]

In parallel to the ALTAI system, the *Future-ai* initiative[37] stands out in the field of health. It includes and develops an ethical evaluation *checklist* system for AI for health, with specific questions and actions covering seven stages of AI development: clinical conceptualisation, requirements gathering, technical

[33] The results in this database are 32 in May 2024: AI Trust Standard & Label; AIShield AI Security Product; Algorithmic Transparency Certification for Artificial Intelligence Systems; Building Trust in Artificial Intelligence; CounterGen; D-Seal; Digital Trust Label; Ethical Problem Solving; Evaluate Library and Evaluation on the Hub (Hugging Face); FRR Quality Mark for (AI Based) Robotics; Giskard; GRACE; Holistic AI Audits; Holistic AI Bias Audits; Holistic AI Governance, Risk and Compliance Platform; Holistic AI Open Source Library; Holistic AI risk mitigation roadmaps; Human-Computer Trust Scale (HCTS); IEEE CertifAIEd; KomplyAi; Model Cards; Naaia; OneTrust AI Governance; Orthrus; SAIF CHECK; Saimple; SECure: A Social and Environmental Certificate for AI Systems; The Certification as a Mechanism for Control of Artificial Intelligence in Europe; The Citrusx Platform; TÜV for Artificial Intelligence; Zupervise; The Certification as a Mechanism for Control of Artificial Intelligence in Europe; The Citrusx Platform; TÜV for Artificial Intelligence; Zupervise

[34] HLEG-European Commission, *Ethical Guidelines for Reliable AI*, 2019, *Ethical Guidelines for Reliable AI*, 2019, https://op.europa.eu/es/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1 in particular Chapter III and listing, pp. 33-41.

[35] https://www.ucc.ie/en/compsci/research/insight/

[36] https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal

[37] https://future-ai.eu/

design, data selection and preparation, AI implementation and optimisation, AI evaluation and deployment, and AI monitoring. It includes elements on fairness, universality, traceability, usability, robustness and explainability, providing examples of mitigation measures to minimise the risks of AI algorithms in healthcare[38]. A comprehensive framework is provided to help developers and clinicians create and evaluate medical AI tools in a systematic way. All of these, in collaboration between researchers, developers and medical professionals to address ethical and technical challenges in medical AI.

Researchers at Oxford University with Floridi have developed *capAI*,[39] a tool designed to perform conformity assessment of AI systems under the AI Act. CapAI provides practical guidelines for converting ethical principles into verifiable criteria, facilitating the ethical design, development, implementation and use of AI. CapAI requirements include risk assessment, data protection, transparency in algorithms and accountability in automated decisions, with a focus on explainability of AI processes and continuous monitoring mechanisms. CapAI is being validated with companies.

The UNESCO Recommendation on the Ethics of Artificial Intelligence of November 2021 is well known.[40] Well, it is worth noting that in 2023, the dissemination of a methodology including qualitative and quantitative indicators grouped into various dimensions that allow a comprehensive assessment of the state of readiness of each country for the ethical implementation of AI.[41] It encompasses the assessment of multiple dimensions: The legal-regulatory dimension assesses the capacity of states to implement regulatory frameworks that ensure data protection, privacy, and gender equality, among other aspects. The social and cultural dimension addresses inclusion, diversity, and public trust in AI, as well as environmental and sustainability policies. The scientific-educational dimension examines the level of AI research and development, including training and educational opportunities. The economic dimension contains the size and strength of the AI ecosystem in the country, including the labour market and investment in AI technologies; the technical and infrastructural dimension looks at the technical infrastructure and connectivity needed for AI development and application. This document also defines the composition of the national assessment team and details how the

[38] A detail of specifications and elements at https://future-ai.eu/checklist/

[39] Floridi, L. et al. *capAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act*, March 2022, http://dx.doi.org/10.2139/ssrn.4064091

[40] https://www.unesco.org/es/artificial-intelligence/recommendation-ethics

[41] UNESCO, *Readiness Assessment Methodology: A Tool for the Recommendation on the Ethics of Artificial Intelligence*, UNESCO, 2023, https://unesdoc.unesco.org/ark:/48223/pf0000385198_spa.

assessment should culminate in a national report and a roadmap for capacity building and improving policy and regulatory frameworks.

ISO/IEC 42001:2023 -Information technology Artificial Intelligence Management system is an international standard developed by ISO/IEC Technical Committee JTC 1/SC 42.[42] This standard, the first of its kind worldwide, provides detailed guidelines for managing risk and security in the development and implementation of AI systems. Aimed at organisations of any size that are involved in the development, implementation, and management of AI systems, it helps to manage AI systems safely and efficiently, meeting the highest standards of quality and transparency. With an international scope, similar to other ISO standards, ISO/IEC 42001 includes guidelines on risk management, information security and compliance with quality standards in the implementation of AI systems. ISO management system standards are recognised globally for their rigorousness and contribution to the continuous improvement of organisations.

Developments from the US NIST should be followed closely. On 29 April 2024, NIST presented a new draft of the *AI Risk Management Framework (AI RMF).*[43] This framework helps organisations to identify, assess, and manage risks associated with AI systems. More than 2500 members participated in the public generative AI working group, highlighting 12 risks and more than 400 actions that developers can implement. In 2024, a *Draft Generative AI Profile* for identifying and managing generative AI risks was released.

From the UK, developments from the *AI Safety Institute*[44] stand out. The AI Safety Institute is the UK's first government-backed organisation dedicated to the safety of Artificial Intelligence during its development phase. In the UK, the *Ada Lovelace Institute* is one of the foremost organisations for Artificial Intelligence and emerging technologies.[45] Its certification model[46] assesses various dimensions of AI systems, including transparency, fairness, privacy and accountability. The certification is aimed at companies and organisations with an initial focus on Europe but is accessible to organisations worldwide. Certification criteria include transparency in algorithms, fairness in outcomes, protection of personal data, accountability in automated decision-making,

[42] https://www.iso.org/standard/81230.html

[43] https://www.nist.gov/itl/ai-risk-management-framework
https://doi.org/10.6028/NIST.AI.100-1

[44] https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute#box-1

[45] https://adalovelaceinstitute.org/

[46] Details at https://www.adalovelaceinstitute.org/wp-content/uploads/2021/12/ADA_Technical-methods-regulatory-inspection_report.pdf

and the ability to explain AI processes. In addition, it requires the implementation of mechanisms for continuous monitoring and evaluation of the ethical impact of the systems.

The *RIAL* initiative, generated by an international team since 2019,[47] encourages the adoption of use restrictions in licences to mitigate risks and harms caused by AI in industry. RIAL licences[48] include behavioural use clauses that restrict and control AI technology applications. Among them, the RIAL Source Code Licence allows code sharing under responsible terms; the RIAL Model Licence sets limitations on the use and distribution of AI models; and the RIAL Data Licence ensures ethical and responsible use of datasets.[49]

The *IEEE CertifAIEd* programme and label help organisations address essential aspects of transparency, accountability, algorithmic bias, and privacy in their AI systems. It sets standards[50] and ethical criteria that include: transparency and values embedded in system design; system accountability and autonomy with learning capabilities; prevention of systematic errors and unwanted behaviour; and privacy protection. In addition, the programme provides an "ecosystem of trainers, evaluators and certifiers".

The *Fairly Trained* Certificate, awarded by a European non-profit organisation, focuses on generative AI models with international scope. It evaluates and certifies Artificial Intelligence products to ensure that their data training models are fairly sourced and respect the rights of creators. The platform provides detailed information on access to the code. The *Fairly Trained* label is awarded to companies that demonstrate the use of ethical and copyright-respectful training data, thereby promoting fairness in AI. Its key requirements focus on ensuring that all data is sourced in a manner that is fair and respectful of the rights of creators.

In Germany, there have been several initiatives. The *German AI Association*,[51] which includes members such as companies and AI experts, is responsible for the *AI Seal of Approval of the KI Bundesverband*. This seal assesses the quality and accountability of AI systems developed by its members. The territorial scope of the seal is focused on Germany. The key requirements of

[47] https://www.licenses.ai/

[48] The theoretical framework can be followed in FAccT of ACM 2022 Behavioural-use Licensing for Responsible AI, and the need for standardisation in On the Standardization of Behavioral Use Clauses and Their Adoption for Responsible Licensing of AI.

[49] https://www.licenses.ai/ai-licenses

[50] Ontological specifications at https://engagestandards.ieee.org/ieeecertifaied.html

[51] https://ki-verband. de/en/projects

the seal[52] include working with criteria set by the association, covering ethics and transparency in AI systems.

The *Towards Auditable AI Systems* certification model[53] is presented as a comprehensive approach to assess and certify AI systems. Developed by the *Fraunhofer Heinrich Hertz Institute* (HHI) together with the *TÜV Association* and *the Federal Office for Information Security* (BSI), the *Towards Auditable AI Systems* model has produced two technical documents: a roadmap in 2021 to examine AI models throughout their lifecycle[54] and a "Certification Readiness Matrix" (CRM) in 2022.[55]

The programme is aimed at developers and auditors of AI systems with an international scope and includes the assessment of documented procedures, training models and implementation practices. Also in Germany, the *AI Cloud Service Compliance Criteria Catalogue* (AIC4)[56] assesses the security and compliance of AI-enabled cloud services. This catalogue, developed by the Federal Office for Information Security (BSI), sets out specific criteria that cloud service providers must meet to ensure that their AI solutions are secure, reliable, and compliant with current regulations. Its scope of application is primarily in Germany, but it can be adopted by international organisations wishing to meet high German security and compliance standards. The portfolio covers risk management, data protection, information security, transparency in AI processes, and compliance with security and privacy regulations.

The (new Danish labelling programme for IT security and responsible use of data)[57] was founded by an independent consortium of stakeholders in Denmark. Its purpose is to assess whether a company meets certain criteria for security and responsible use of data, with the number of criteria varying according to the desired label level.[58] The main objective of the programme is to make visible whether a company has good IT security and a responsi-

[52] Algún detalle en https://www.am.ai/171feadfe65186a2f4d42891383a58d7/KIBV_Guetesiegel_190302_o.pdf

[53] https://www.hhi.fraunhofer.de/en/departments/ai/technologies-and-solutions/auditing-and-certification-of-ai-systems.html

[54] BSI, "Towards Auditable AI Systems: Current status and future directions" May 2021. Also, BSI, *Towards Auditable AI Systems AI Cloud Service*, https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems_2022.pdf?__blob=publicationFile&v=4

[55] "Towards Auditable AI Systems: From Principles to Practice" of May 2022.

[56] BSI, *Compliance Criteria Catalogue (AIC4)*, https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/CloudComputing/AIC4/AI-Cloud-Service-Compliance-Criteria-Catalogue_AIC4.pdf?__blob=publicationFile&v=4

[57] https://d-seal.eu/

[58] Some detail on https://d-seal.eu/criteria/

ble use of data. It is aimed at citizens and consumers in Denmark in their relationship with companies, covering a national territorial scope. The key requirements of the programme include a number of checks per criterion that varies and is regularly updated. In addition, companies must register and apply for the application of the label.

Finally, among other initiatives, the *Responsible Artificial Intelligence Institute* (RAI Institute) offers a Responsible AI certification programme.[59] Initially, this programme focuses on AI systems developed in North America, but with a potentially global scope. The[60] certification assesses six main dimensions: system operations, explainability and interpretability, accountability, consumer protection, fairness and absence of bias, and robustness. The certification process includes a review, development of the implementation framework, evaluation testing and adjustment, training, and calibration.

V. In conclusion

This study has analysed AIA regulation with respect to AI systems that, essentially, does not regulate, i.e., systems that are not high-risk. In any case, and as a starting point, these systems are subject to other applicable regulations, such as product safety or data protection. While the AIA focuses on high-risk AI systems and imposes strict obligations, it also encourages the creation and development of an ecosystem of codes of conduct, seals, and certification schemes in the field of AI within the EU for non-high-risk systems. It should be noted that this is in line with the rest of the world, where self-regulatory codes and regulatory systems that are softer than AIA, such as the Hiroshima Code agreed in 2023 by the G7, are prioritised in the US or the UK. In the coming years, certifications of varying scope, origin, sector, nature, and intensity of requirements will be developed. Time and the market will determine the usefulness and success of these voluntary instruments for non-high risk systems. These codes, seals, and certifications will be a strong underpinning to ensure the quality and safety of systems, their compliance with fundamental ethical principles such as transparency, accountability and the prevention of algorithmic bias. In addition, as the AIA notes, they can play an important role in sustainability, inclusiveness, and generally building user and consumer confidence in AI technologies.

[59] https://www.responsible.ai/how-we-help/#certification

[60] Details at https://www.responsible.ai/wp-content/uploads/2024/02/RIAI-Certification-Guidebook.pdf

Having set out the scope of Article 95 of the AIA, it has been pointed out that the failure to regulate the general principles proposed by the EU Parliament was a missed opportunity. These principles would have applied to all AI systems, not just high-risk ones. The important role played by the principles in Article 5 of the GDPR is well known. Normative recognition of these principles, already well known in the field of AI ethics, could have had great potential as legal principles and rules applicable to all types of AI.

As for the development of AI labels and certifications, this is still an incipient field on which the Spanish government bet early in its 2020 ENIA. Perhaps too early, to the point that the last ENIA of May 2024 seems to have forgotten the initiative of a Spanish AI Seal that was planned and, at least in theory, well planned for implementation.

The study has analysed more than thirty AI certification initiatives and tools, both at European and international level. Among them, more than a dozen of the most relevant or well-known ones have been described. The review shows the diversity and richness of approaches available to address voluntary AI certification models for non-high-risk systems.

These AI certification schemes and codes of conduct are likely to remain in the shadow of the *hard* obligations for high-risk systems set out in the AIA and need to be developed with more concrete criteria, harmonised standards, and technical specifications. However, in the future, it will be essential to continue to develop and refine this ecosystem of voluntary certifications, and it is possible that they will eventually become a strong and effective presence. It will therefore be important to foster collaboration between governments, private organisations, and civil society to ensure that AI systems are developed and deployed to make responsible and ethical AI, which the EU advocates in its AIA, more effective, even for systems that are not high-risk.

ARTICLE 50 OF THE AI ACT AND THE TRANSPARENCY OBLIGATIONS FOR PROVIDERS AND DEPLOYERS OF CERTAIN ARTIFICIAL INTELLIGENCE SYSTEMS

Agustí Cerrillo i Martínez
Professor of Administrative Law at the Universitat Oberta de Catalunya

I. Limited-risk Artificial Intelligence systems

The AIA classifies Artificial Intelligence (AI) systems according to the risk they may pose to public interests and fundamental rights protected by EU law. Indeed, as recital 26 states: "In order to introduce a proportionate and effective set of binding rules for AI systems, a clearly defined risk-based approach should be followed. That approach should tailor the type and content of such rules to the intensity and scope of the risks that AI systems can generate". Based on this approach, the AIA prohibits the use of certain AI systems (Article 5 AIA) and classifies others as high-risk systems (Article 6 AIA) because of their impact on EU public interests or fundamental rights.

In addition, the AIA warns that certain AI systems intended to interact with natural persons or to generate content may generate other specific risks such as impersonation, deception or manipulation of persons. As Peguera warns, this is not strictly speaking a specific category of risk[1], although the AIA foresees that users or recipients of the results of these AI systems must be able to be aware that they are dealing with AI systems or that the results obtained have been artificially generated. In this regard, Article 50 AIA provides for various transparency obligations, which will be analysed in this chapter. The aim is to ensure that any person who comes into contact with these systems or the results they generate, can be aware of these circumstances, make informed decisions, or avoid a given situation.

The following pages first outline the evolution of the regulation of the transparency obligations of certain AI systems from the proposal made by the European Commission in 2021 to the text finally published in the OJEU. It then describes the different AI systems concerned and analyses the transparency obligations envisaged for each of them. Finally, it concludes with some final reflections.

[1] Peguera Poch, M., *La propuesta de Reglamento de AI: Una intervención legislativa insoslable en contexto de incertidumbre*, in Peguera Poch, M., Perspectivas regulatorias de la inteligencia artificial en la Unión Europea, Reus, Madrid, 2023.

II. Development, processing and final content of Article 50 the AIA

The provision of transparency obligations for certain AI systems has been foreseen in the AIA since the proposal presented by the European Commission in 2021 (COM(2021) 206 final).

Indeed, even in that first text, Title IV was included, consisting only of Article 52, which explicitly provided for three obligations whose regulation could not affect the transparency obligations generally provided for in the regulation of high-risk systems (Title III).

According to the explanation accompanying the proposal, Article 52 focused on AI systems that could give rise to specific risks of manipulation and therefore provided for specific transparency obligations in the form of an obligation to disclose this circumstance in order to enable the person concerned to make informed decisions or to avoid a certain situation.

First, there was an obligation for providers to ensure that AI systems intended to interact with natural persons are designed in such a way that persons can be informed that they are interacting with an AI system. Secondly, it included an obligation for users of emotion recognition or biometric categorisation systems to inform natural persons exposed to them about their operation. Again, in this case, the proposal incorporated some exceptions. Thirdly, an obligation for users of certain systems that generate or manipulate the content of images, sounds, or videos that could mislead people into believing that they are genuine or truthful that they have been artificially generated or manipulated, was foreseen. Finally, in all three cases the obligations were established with some limitations and exceptions.

One of the main innovations introduced by the Council was in relation to the exceptions to the first obligation[2]. The Commission proposal provided for a limitation of the reporting obligation "in situations where this is obvious from the circumstances and context of use". Instead, the Council made it more specific by providing that "in situations where this is obvious from the point of view of a reasonably well-informed, observant and circumspect legal person, taking into account the circumstances and the context of use". This was the wording finally adopted. In relation to the exceptions to the first obligation, the Council also incorporated that the systems covered by it -for the purpose of detection, prevention, investigation or prosecution of criminal offences- should operate subject to appropriate safeguards for the

[2] According to the document approved on 25 November 2022. Accessible at: https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/es/pdf (last accessed February 2024).

rights and freedoms of third parties. In relation to the second obligation, the Council proposed to distinguish in two paragraphs the regulation relating to emotion recognition systems and biometric profiling systems, although, in general terms, the scope of the obligation was the same as that already foreseen in the Commission proposal in relation to each of these systems, adding only, as in the case of the exception to the first obligation, that the rights and freedoms of third parties should be respected. Finally, with regard to the third obligation, the Council only proposed a drafting change in relation to the scope of the right to freedom of expression. Also generally, in relation to all three obligations, it was proposed by the Council that information should be provided in a 'clear and conspicuous manner at the latest on the occasion of the first interaction or exhibition'. Finally, it was also suggested to include in the regulation that the obligations foreseen in Article 52 would not only not affect the provisions of Title III as already foreseen in the Commission proposal, but would also not affect "other transparency obligations for users of AI systems laid down in Union or national law".

The European Parliament also tabled a number of amendments to the Commission's proposal[3]. In relation to the first obligation, it proposed to specify that information should be provided "in a clear, intelligible and timely manner". Furthermore, the Parliament's amendments suggested that "this information shall also disclose which functions are enabled by AI, whether there is human surveillance and who is responsible for the decision-making process, as well as the existing rights and processes under Union and national law that allow natural persons or their representatives to object to the application of such systems and to seek judicial redress against decisions taken by or damage caused by AI systems, including their right to request an explanation" (amendment 484). As regards the second obligation, the European Parliament also proposed that the information should be timely, clear and intelligible and that, in the case of processing of biometric data, the consent of the natural person exposed to it should be obtained (amendment 485). As regards the third obligation, the European Parliament suggested providing for an obligation to inform, where possible, the natural or legal person who generated or handled the content. It was also proposed that inauthentic content should have to be labelled, in accordance with the state of the art and relevant harmonised standards and specifications, in order to be clearly visible (amendment 486). Furthermore, the European Parliament proposed that the third obligation should not be required where the generation or manipula-

[3] According to text adopted on 14 June 2023. Accessible at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_ES.html (last accessed February 2024).

tion is authorised by law or, where the content is part of a clearly creative, satirical, artistic or fictional cinematographic work, images from video games and similar works or formats, does not hinder the presentation of the work (amendment 487). Finally, it was proposed that the information should be made available to natural persons in an accessible form on the occasion of the first interaction or exhibition (amendment 488).

As we will have the opportunity to analyse in detail in the following sections, the text finally adopted -which is finally Article 50 AIA- has maintained the spirit of the Commission's proposal, but has included most of the proposals and amendments made. Mainly those made by the Council. However, perhaps the main novelty incorporated during the procedure is the decision to regulate general purpose Artificial Intelligence systems (GPAI), which has resulted in the inclusion of a specific transparency obligation for those systems that are capable of generating audio, images, videos or synthetic texts to report on them.

III. The scope of the transparency obligations provided for in Article 50 AIA

Article 50 AIA regulates various transparency obligations that providers and deployers of certain AI systems must comply with. These obligations apply to a limited set of AI systems whose use may bring many benefits but may also entail some risks that can potentially have a wide-ranging impact.

As we will see in the following pages, these AI systems do not inherently or directly or exclusively generate risks against public interests or with respect to health, security or fundamental rights. But they can be used in ways that have a negative impact on society by facilitating disinformation, manipulation, fraud, deception or simply confusion.

The AIA has opted not to prohibit or restrict their use but to warn their users to be aware about the use of these AI systems. In relation to this legislative option, some authors have questioned whether it will be sufficiently adequate to avoid a negative impact on public interests or a violation of fundamental rights. They have therefore suggested that the AIA should have provided for stricter regulation, e.g., a conformity or fundamental rights impact assessment[4]. It may be too early to tell and it will be necessary to assess the

[4] Barkane, I., "*Questioning the EU proposal for an Artificial Intelligence Act: The need for prohibitions and a stricter approach to biometric surveillance 1*", *Information Polity*, No 27 (2022).

effectiveness of this option and, if necessary, update the regulation to achieve the intended purposes.

The AIA has foreseen that the transparency obligations applicable to AI systems described below do not affect the possible application of the requirements and obligations foreseen in the rule itself for high-risk AI systems (Article 50.6). It has also recognised that they do not affect other transparency obligations that may be envisaged by Member States or by the European Union itself (Article 50.6). For example, those that may derive from transparency legislation when AI systems are used by public administrations.

Finally, as a general rule and applicable to the different cases under analysis, Article 50.5 of the AIA provides that the information given by providers and deployers in compliance with transparency obligations must be accessible and comprehensible. For this reason, it must be provided in a clear and distinguishable manner. It must also be timely. Thus, it needs to be provided at the latest on the occasion of the first interaction or exposure. Finally, the information provided must comply with the applicable accessibility requirements provided for in Directive (EU) 2016/2102 of the European Parliament and of the Council of 26 October 2016 on the accessibility of websites and mobile applications of public sector bodies and Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on accessibility requirements for products and services.

IV. Artificial Intelligence systems interacting directly with natural persons

1. Covered systems

The first paragraph of Article 50 regulates transparency obligations for AI systems that interact directly with natural persons. These are AI systems that enable people to interact with devices in natural language, spoken or written, in such a way that they are able to understand the content of the message and act accordingly.[5]

Throughout its articles, the AIA does not include any definition of these systems, nor does it determine the characteristics they should have.

Perhaps the most widespread of the AI systems that interact with people are the conversational robots or chatbots used by many companies and public

[5] Cerrillo I Martínez, A., "*Robots, virtual assistants and automation of public administrations*", *Revista Galega de Administración Pública*, núm 61 (2021).

administrations and the virtual assistants embedded in different devices[6]. But beyond this, there are other applications of these AI systems, for example for the remote control of aerospace devices or underwater vehicles.[7]

As the quality of these AI systems increases, people interacting with them find it more difficult to know whether they are interacting with a person or a machine. At the same time, concern is growing among them to the point of generating different types of rejection of their use.[8]

2. Scope of obligations

In its first paragraph, Article 50 AIA establishes an obligation to design and develop these systems in such a way as to provide information on the fact that the natural person may know that he or she is interacting with an AI system.

The obliged party is the providers of AI systems, who must design them in such a way that the obligation to provide information can be fulfilled. As Veale and Zuiderveen Borgesius point out, it might have been desirable for the AIA to refer not only to providers but also to those responsible for deployment to ensure that in any case the information reaches the person interacting with the AI systems[9]. In particular for those cases where the system is integrated into another service that is ultimately received by the user.

The AIA does not indicate how the system should be designed or how the information should be provided, with the system provider determining this as long as the intended purpose is achieved. This lack of criteria may lead to each provider determining the scope of the information it supplies, which may have a negative impact on transparency.[10]

The obligation is limited in cases where it is obvious to a reasonably well-informed, attentive and circumspect natural person, taking into account the circumstances and context of use, that he or she is interacting with an AI system. The qualifiers introduced in the regulation during its passage under-

[6] Adamopoulou, E. and Moussiades, L., "*Chatbots: History, technology, and applications*", *Machine Learning with Applications*, no. 2 (2020).

[7] Sheridan, T. B., "*Human-robot interaction: status and challenges*", *Human factors*, No 58 (2016).

[8] Bartneck, C., Belpaeme, T., Eyssel, F., Kanda, T., Keijsers, M. and Šabanović, S., *Human-robot interaction: An introduction*, Cambridge University Press, Cambridge, 2020.

[9] Veale, M. and Zuiderveen Borgesius, F., "*Demystifying the Draft EU Artificial Intelligence Act-Analysing the good, the bad, and the unclear elements of the proposed approach*", *Computer Law Review International*, no. 22 (2021).

[10] Stuurman, K. and Lachaud, E., "*Regulating AI. A label to complete the proposed Act on Artificial Intelligence*", *Computer Law & Security Review*, No 44 (2022).

line the desire to ensure that providers of AI systems are particularly careful to foresee how the information can actually reach the individuals concerned and do not end up shifting the responsibility for locating or obtaining the information to them.

On the other hand, the obligation in paragraph 1 is exempted for systems authorised by law to detect, prevent, investigate or prosecute criminal offences unless such systems are available for the public to report a criminal offence. In this respect it cannot be ignored that, as already noted in the Communication *Building trust and confidence in human-centric Artificial Intelligence* [COM(2019) 168 final] of 8 April 2019, “AI can also help detect fraud and cybersecurity threats and enable law enforcement agencies to fight crime more effectively”. However, the Commission itself has also warned how “the known use of similar technologies for surveillance purposes, by public or private companies, may raise concerns and reduce trust in the digital economy among individuals and organisations” (Communication *Towards a thriving data economy* COM(2014) 442 final of 2 July). In any case, this exception, which is also foreseen in relation to other AI systems covered by Article 50, should be subject to appropriate safeguards for the rights and freedoms of third parties.

V. Artificial Intelligence systems that generate synthetic content

1. Covered systems

AI systems have evolved rapidly in recent times to acquire, among other capabilities, the ability to generate synthetic content, in other words, artificially generated content. This content is so realistic that a person would not be able to tell that it was created by an AI system.

Synthetic content can be of different types. In particular, Article 50.2 refers to AI systems generating synthetic audio, image, video or text content.

These AI systems that create synthetic content specifically include general-purpose AI systems. These AI systems have great potential, but also pose numerous risks that have been carried over into the many discussions that have taken place throughout the AIA pipeline. Indeed, as the weekly Politico headlined in March 2023 “ChatGPT broke EU plan to regulate Artificial Intelligence”[11]. In fact, the European Commission’s proposal did not refer

[11] Accessible at: https://www.politico.eu/article/eu-plan-regulate-chatgpt-openai-artificial-intelligence-act/ (last accessed March 2024).

to general-purpose Artificial Intelligence systems[12]. Indeed, this proposal focused on conventional AI models[13]. It was not until the Slovenian Presidency in 2021 that a first mention was included, which was subsequently deepened by the French Presidency.

Apart from the analysis in other chapters, it is worth mentioning at this point that these AI systems are very complex, are trained with millions of data and are made up of millions of parameters. But their main characteristic is that they can perform very different tasks, some of them not initially foreseen. To do so, they consume large volumes of computing power and therefore also a lot of energy.[14]

Despite the quality of the results that can be obtained and the uses that can be made of these AI systems -for example, in the field of medicine[15], urban planning[16], education[17], or even in public administration-[18], they are still texts, images or videos generated from a combination of information based on probabilities. As has been shown, these systems are not able to understand the generated content (what has been baptised with the metaphor of the *stochastic parrot*)[19]. Furthermore, these models can be *delusional*, i.e., they provide very credible or convincing results that do not correspond to the algorithm's training data and can therefore be false[20]. Moreover, these AI systems do not

[12] Moreira, N. A., Freitas, P. M. and Novais, P., *The AI Act Meets General Purpose AI: The Good, The Bad and The Uncertain*, in Moniz, N., Vale, Z., Cascalho, J., Silva, C. and Sebastião, R., EPIA Conference on Artificial Intelligence, Springer, Cham, 2023.

[13] Hacker, P., Engel, A. and Mauer, M., *Regulating ChatGPT and other large generative AI models*, 2023.

[14] OECD, *Measuring the environmental impacts of Artificial Intelligence compute and applications*, 2022.

[15] Sallam, M., *ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns*, MDPI, 2023.

[16] Wang, D., Lu, C.-T. and Fu, Y., "*Towards automated urban planning: When generative and chatgpt-like ai meets urban planning*", *arXiv preprint arXiv:2304.03892*, no. (2023).

[17] Grassini, S., '*Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings*', *Education Sciences*, No 13 (2023).

[18] Huang, J. and Huang, K., *ChatGPT in Government*, in Huang, K., Wang, Y., Zhu, F., Chen, X. and Xing, C., Beyond AI: ChatGPT, Web3, and the Business Landscape of Tomorrow, Springer Nature Switzerland, Cham, 2023.

[19] Bender, E. M., Gebru, T., Mcmillan-Major, A. and Shmitchell, S., *On the dangers of stochastic parrots: Can language models be too big?*, 2021; Srivastava, V., "*When Stochastic Parrots Learn to Swim: The Regulation of General Purpose Artificial Intelligence in the EU*", núm (2023).

[20] Triguero, I., Molina, D., Poyatos, J., Del Ser, J. and Herrera, F., "*General Purpose Artificial Intelligence Systems (GPAIS): Properties, definition, taxonomy, societal implications and responsible governance*", *Information Fusion*, núm 103 (2024).

escape the risks that AI in general entails in terms of data quality[21]. Finally, these systems may infringe intellectual property rights.[22]

In addition to these risks, as we will see later, the concern generated around these AI systems lies in the fact that they can be used to generate content that, due to their verisimilitude or hyperrealism, can lead to manipulation or disinformation.[23]

It is because of the existence of all these risks, but also because of the possibility of new ones emerging as uses evolve -the *black swans damages* referred to by Kolt[24]- that the AIA has provided for transparency obligations in respect of these AI systems.

2. Scope of obligations

Transparency obligations for AI systems generating synthetic content are addressed to providers. Again in this case, it has been suggested that it would have been appropriate to provide for some obligation on the deployers as well[25]. Indeed, while generally in cases where the deployer modifies the intended purpose of an AI system that has already been placed on the market or put into service in such a way that it becomes a high-risk AI system (Article 25.1.c AIA), it should be considered as the provider, in other cases the modification may not be so substantial but nevertheless have an impact on how the result finally reaches the end-user.

AI system providers must ensure that the generated result (output information from the AI system) is marked and that it is possible to detect that the result has been artificially generated or manipulated.

The mark must be machine-readable, i.e., it must be in a structured file format that allows software applications to easily identify, recognise and extract specific data, including factual statements and their internal structure[26]. To this end, the paragraph itself foresees that providers should use technical

[21] Moreira, N. A., Freitas, P. M. and Novais, P., *The AI Act Meets General Purpose AI: The Good, The Bad and The Uncertain*, op.cit.

[22] Lucchi, N., "*ChatGPT: a case study on copyright challenges for generative Artificial Intelligence systems*", *European Journal of Risk Regulation*, No. (2023).

[23] Hacker, P., Engel, A. and Mauer, M., *Regulating ChatGPT and other large generative AI models*, 2023.

[24] Kolt, N., "*Algorithmic black swans*", *Washington University Law Review*, No 101 (2023).

[25] Edwards, L., *Regulating AI in Europe: four problems and four solutions*, Ada Lovelace Institute, London, 2022.

[26] Article 2.13 Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information.

solutions that are efficient, interoperable, robust and reliable as far as technically feasible. In defining the technical solutions to be used, they should also take into account the specificities and limitations inherent in each type of content created (audio, image, video or text), the costs of implementation and the generally recognised state of the art, as reflected in the relevant technical standards.

To this end, Article 50.7 provides that the AI Office shall encourage and facilitate the drawing up of codes of practice at Union level to facilitate the effective implementation of this obligation. The Commission is also empowered to adopt implementing acts to approve these codes of practice in accordance with Article 56 AIA and, if it considers that this is not appropriate, to adopt an implementing act specifying the common rules for the implementation of these obligations in accordance with Article 98 AIA.

The information provided must be sufficiently clear so that the recipient of the result can be aware that the content has been artificially generated or manipulated.

Article 50.2 of the AIA limits the obligation to cases where the AI systems perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof. In these circumstances, AI system providers should not ensure that the output information is marked up or that it can be detected as artificially generated or manipulated.

Finally, as in the case of AI systems that interact directly with individuals, systems that generate content are exempted from the transparency obligation when they are authorized by law to detect, prevent, investigate, or prosecute criminal offenses.

VI. Artificial Intelligence systems for emotion recognition

1. Covered systems

In recent decades, progress has been made in the development of AI systems that are capable of automatically detecting an inherent element of people such as emotions.

Emotion recognition AI systems are a type of so-called affective computing, i.e., computers that have various capabilities related to emotions such as recognition, expression, modelling, communication or reaction[27]. Emotion

[27] Picard, R. W., *Affective computing*, MIT press, Boston, 2000.

recognition AI systems aim to enable machines to measure, assess, predict or react to people's emotional states based on various data extracted from physical or physiological elements that may be in texts, voices, images or videos or captured through biometric sensors[28]. Thus, emotion recognition AI systems are able to convert emotions into data[29]. However, they do not allow computers to feel or express emotions themselves and are therefore defined as a type of weak AI.[30]

The AIA defines emotion recognition AI systems as systems for the purpose of identifying or inferring emotions or intentions of natural persons on the basis of their biometric data (Article 3.39). The AIA refers to emotion recognition AI systems and not simply to detection systems. Thus, emotion detection AI systems that do not involve recognition will fall outside the transparency obligation under Article 50.

These AI systems aim to associate a facial expression, speech cadence or body movement with a certain emotion (e.g., fear, sadness, anger, joy, surprise or disgust). In this direction, AIA refers to emotions or intentions such as happiness, sadness, indignation, surprise, disgust, distress, enthusiasm, embarrassment, contempt, satisfaction and amusement. On the other hand, it excludes physical states (e.g., pain or tiredness) (recital 18).

This is done by examining physical signals (such as facial expressions, eye or body movement, speech or text, or body postures) or physiological signals (e.g. body temperature, heart rate or breathing rate) captured by different sensors.

Today, numerous applications are already being made of these AI systems in the fields of health (e.g., to detect pain suffered by a patient) and mental health (e.g., to identify mood). Business and commercial uses have also been expanding (e.g., in applications that recommend products based on customers' moods). Automated emotion recognition is also being used in education (e.g., for personalisation of learning or identification of learning difficulties). Similarly, in public security, projects such as Avatar -the Automated Virtual Agent for Real-Time Truth Assessment developed in the US for border con-

[28] Mcstay, A., "*Emotional AI and EdTech: serving the public good?*", *Learning, Media and Technology*, No 45 (2020).; Gremsl, T. and Hödl, E., "*Emotional AI: Legal and ethical challenges 1*", *Information Polity*, No 27 (2022); Podoletz, L., "*We have to talk about emotional AI and crime*", *AI & SOCIETY*, No 38 (2023).

[29] Steindl, E., "*Does the European Data Protection Framework Adequately Protect Our Emotions? Emotion Tech in Light of the Draft AI Act and Its Interplay with the GDPR*", *Eur. Data Prot. L. Rev.*, No. 8 (2022).

[30] Mcstay, A., "*Emotional AI and EdTech: serving the public good?*", *Learning, Media and Technology*, No 45 (2020).

trol- which analyses the verbal and non-verbal language of travellers seeking to enter the country[31]; or iBorderCtrl, the border control project funded by the European Commission[32] have been promoted. Finally, these AI systems are also used in emotion recognition in other applications such as those incorporated in some vehicles to detect if a driver at the wheel is falling asleep,[33] or those used by some companies to monitor and control the activity carried out in the workplace[34]. However, in relation to the latter applications, it should be borne in mind that recital 18 AIA does not consider systems used to detect fatigue in drivers or professional drivers in order to avoid accidents to be included among emotion recognition systems.

The extent of devices that have some form of emotion recognition AI system is large, estimated to extend to 10% of devices and predicted to reach a value of $37 billion by 2026.[35]

Despite the advances that have been made, it cannot be ignored that there is no academic consensus on the relationship between emotions and their physical or physiological expression[36]. In recent years, different voices have stated that emotion recognition is not scientifically proven[37]. They have also pointed out that expressions of emotions, for example through the face, are not the same depending on the context or culture[38]. Thus, they have stressed that these AI systems often fail to achieve the expected results.[39]

The impact that the use of these systems may have on fundamental rights cannot be underestimated either. In fact, certain emotion recognition AI systems are considered in the AIA as high-risk systems (section 1.c Annex III).

In particular, the use of emotion recognition AI systems may have an

[31] Cotino Hueso, L., "*Artificial Intelligence systems with facial recognition and biometric data. Mejor regular bien que prohibir mal*", *El Cronista del Estado Social y Democrático de Derecho*, núm 100 (2022).

[32] Romano, A., "*Drets fonamentals i intel- ligència artificial emocional en iBorderCtrl: reptes de l'automatització en l'àmbit migratori*", *Revista Catalana de Dret Públic*, núm (2023).

[33] Mcstay, A. and Urquhart, L., '*In cars (are we really safest of all?): interior sensing and emotional opacity*', *International Review of Law, Computers & Technology*, No 36 (2022).

[34] Kumar, M., Aijaz, A., Chattar, O., Shukla, J. and Mutharaju, R., "*Opacity, Transparency, and the Ethics of Affective Computing*", no (2024).

[35] Crawford, K., "*Time to regulate AI that interprets human emotions*", *Nature*, no. 592 (2021).

[36] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M. and Pollak, S. D., "*Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements*", *Psychological science in the public interest*, no. 20 (2019).

[37] Barkane, I., "*Questioning the EU proposal for an Artificial Intelligence Act: The need for prohibitions and a stricter approach to biometric surveillance 1*", *Information Polity*, No 27 (2022).

[38] Heaven, D., "*Why faces don't always tell the truth about feelings*", *Nature*, No 578 (2020).

[39] Katirai, A., "*Ethical considerations in emotion recognition technologies: a review of the literature*", *AI and Ethics*, No. (2023).

impact on privacy and personal data protection[40]. Indeed, we cannot ignore the sensitivity of *emotional data*, which in certain cases can be considered as personal data to the extent that they can be used to identify a person[41]. Even emotional data can also be categorised as biometric data that require special protection[42]. Although the GDPR does not explicitly refer to this, it is clear that in view of the definition of biometric data in Article 4.14, emotion data may in many cases be considered as biometric data with the consequences that may arise from this, such as the prohibition of processing if none of the cases provided for in Article 9.2 of the GDPR apply.

In fact, the definition of the emotion recognition AI system is linked to biometric data. Article 3.34 AIA -as well as Article 4.14 GDPR- defines biometric data as personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, such as facial images or dactyloscopic data, and therefore emotion data may amount to biometric data. Therefore, to the extent that emotion recognition AI systems involve processing of personal data they will not only fall within the scope of Article 50 AIA but also the provisions of the GDPR or, where applicable, Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of individuals with regard to the processing of personal data by the institutions, bodies, offices and agencies of the Union and on the free movement of such data or Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data.

Moreover, to the extent that the emotion data processed by the AI system can be considered as biometric data, these AI systems will also be considered as high-risk AI systems as set out in Annex III AIA by reference to Article 6.2 AIA. It is precisely for this reason that the AIA has been criticised for describing emotion recognition AI systems as limited-risk systems, considering it insufficient to address the risks they may eventually generate and to

[40] Podoletz, L., "*We have to talk about emotional AI and crime*", *AI & SOCIETY*, no 38 (2023).

[41] Gremsl, T. and Hödl, E., "*Emotional AI: Legal and ethical challenges 1*", *Information Polity*, No 27 (2022).

[42] Romano, A., "*Drets fonamentals i intel- ligència artificial emocional en iBorderCtrl: reptes de l'automatització en l'àmbit migratori*", *Revista Catalana de Dret Públic*, num (2023).

adequately inform users of the impact they may have or the intrusion they may pose.[43]

In any case, as indicated above, the application of the transparency obligations under Article 50 shall not affect the requirements and obligations applicable to high-risk schemes.

2. Scope of obligations

In recent years, there has been a growing fear of *emotive surveillance*[44], which has led to proposals from some authorities to regulate its use[45], and, until then, to ban it.[46]

The AIA, beyond their possible categorisation as high-risk AI systems, has opted in Article 50.3 to establish transparency obligations for emotion recognition AI systems to be complied with by those responsible for their deployment.

Unlike the AI systems discussed in the previous sections, in this case it is not the providers of the AI systems but their users, who are responsible for the deployment, who will have to report on the performance of the system.

Indeed, according to Article 50.3, deployers shall inform natural persons exposed to the emotion recognition AI system. This is intended to ensure that the persons concerned can easily be made aware of the use of these systems and that their emotions can be automatically recognised through AI.

The information should relate to the functioning of the system, i.e., in accordance with Article 3.18 AIA, the ability of the system to achieve its intended purpose, i.e., the recognition of emotions. This should not only enable the person concerned to be aware of the existence of the AI system, but also to decide whether he wants to be subject to automated emotion recognition or the results or consequences he may have.

The information provided should address the intended use of the AI system, its specific context and conditions of use. In doing so, the deployer should take into account the information provided by the provider in the in-

[43] Steindl, E., "*Does the European Data Protection Framework Adequately Protect Our Emotions? Emotion Tech in Light of the Draft AI Act and Its Interplay with the GDPR", Eur. Data Prot. L. Rev.*, no. 8 (2022); Veale, M. and Zuiderveen Borgesius, F., "*Demystifying the Draft EU Artificial Intelligence Act-Analysing the good, the bad, and the unclear elements of the proposed approach", Computer Law Review International,* No. 22 (2021).

[44] Steindl, E., "*Does the European Data Protection Framework Adequately Protect Our Emotions?" op.cit.*

[45] Crawford, K., "*Time to regulate AI that interprets human emotions", Nature*, no. 592 (2021).

[46] AI Now Institute, *2019 Report*, 2019.

structions for use, promotional and sales materials and statements, and technical documentation (Article 3.12 AIA).

In specifying the scope of this information, consideration could be given to the provisions of Article 13 for high-risk systems requiring that the information to be provided be "concise, complete, correct and clear information that is relevant, accessible and comprehensible" on aspects such as the identity and contact details of the provider; the characteristics, capabilities and limitations of performance of the AI system (intended purpose; level of accuracy; any known or foreseeable circumstances that could give rise to risks to health and safety or fundamental rights; the performance of the system in relation to the persons to whom the system is to be used; information to enable the results of the system to be interpreted and used appropriately); changes made at the time of the initial conformity assessment; monitoring measures for the system to be used; information to enable the results of the system to be interpreted and used appropriately); performance in relation to the persons for whom the system is to be used; information to enable the results of the system to be interpreted and used appropriately); changes made at the time of the initial conformity assessment; human supervision measures; required hardware and software resources, expected lifetime and required maintenance and care measures; a description of the mechanisms included in the AI system to enable users to collect, store and interpret records appropriately.

Beyond the content, we must remember the need for information to reach the people concerned in an appropriate manner and to be provided in a clear and distinguishable way.

Finally, as in other cases provided for in Article 50 AIA, this paragraph provides for an exception to the obligation of transparency in cases of AI systems which are permitted by law to detect, prevent and investigate criminal offences provided that adequate safeguards are in place for the rights and freedoms of third parties, and in accordance with Union law.

VII. Artificial Intelligence systems used for biometric categorisation

1. Covered systems

Biometrics is one of the most widespread applications of AI systems that may pose the greatest risks to the rights of individuals.

Biometric data are generally used to establish or authenticate a person's identity on the basis of biological elements (e.g. iris, face, fingerprints or DNA), behavioural elements (e.g., gait or voice) or even acquired elements (e.g., marks, tattoos)[47]. But they can also be used to profile or classify people into groups[48]. This is recognised by the AIA when it states that "biometric data may allow the authentication, identification, or categorisation of natural persons and for the recognition of emotions of natural persons" (Recital 14).

According to the AIA, biometric categorisation systems are "an AI system for the purpose of assigning natural persons to specific categories on the basis of their biometric data, unless it is ancillary to another commercial service and strictly necessary for objective technical reasons" (Article 3.40).

AI systems included in this section should only serve the purpose of categorisation. However, it is not possible to identify a person (find out who they are by comparing their biometric data to biometric data of people stored in a database) or verify their identity (confirm their identity by comparing their biometric data to biometric data that has already been provided). Nevertheless, as the European Parliament, among others, has warned, the distinction between biometric identification systems and biometric categorisation systems may be arbitrary in that categorisation may use data that may eventually allow for identification.[49]

Indeed, this distinction is important because in view of the AIA, systems using biometric data can be classified into three different categories with very different regulations[50]. In particular, it is considered a prohibited AI practice "the placing on the market, the putting into service for this specific purpose, or the use of biometric categorisation systems that categorise individually natural persons based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation; this prohibition does not cover any labelling or filtering of lawfully acquired biometric datasets, such as images, based on biometric data or categorizing of biometric data in the area of law enforcement" (Article 5.1.g AIA).

Secondly, 'AI systems intended to be used for biometric categorisation according to sensitive or protected attributes or characteristics based on the inference of those attributes or characteristics' (Annex III) are included among high-risk AI systems, to the extent that their use is permitted by applicable Union or national law.

[47] De Keyser, A., Bart, Y., Gu, X., Liu, S. Q., Robinson, S. G. and Kannan, P., "*Opportunities and challenges of using biometrics for business: Developing a research agenda"*, *Journal of Business Research*, No 136 (2021).

[48] Mobilio, G., "*Your face is not new to me-Regulating the surveillance power of facial recognition technologies"*, *Internet Policy Review*, no. 12 (2023).

[49] European Parliament, *Regulating facial recognition in the EU*, 2021.

[50] Barkane, I., "*Questioning the EU proposal for an Artificial Intelligence Act: The need for prohibitions and a stricter approach to biometric surveillance 1"*, *Information Polity*, No 27 (2022).

Finally, systems using biometric data can be considered as limited risk systems if they are only used for biometric categorisation. In this case, the decision is based on the lesser impact such systems may have on fundamental rights.[51]

Furthermore, the use of biometric data may determine the applicable legal regime. In particular, if emotion-related data do not allow the unique identification of the individual, they are neither personal data nor specially protected data (Articles 14.14 and 9 GDPR). Otherwise, they will be and, as follows from Article 50.3 AIA, those responsible for deploying a biometric categorisation system must treat them in accordance with the GDPR, Regulation (EU) 2018/1725 and Directive (EU) 2016/680, as applicable.

Biometric categorisation is being used in the commercial field to find out consumer preferences or to personalise marketing actions. It is also being used in the human resources departments of companies during the selection process.

2. Scope of obligations

Finally, in the AIA, the regulation of the transparency obligation for biometric categorisation systems has been carried out jointly with that of emotion recognition AI systems.

For biometric categorisation AI systems, obliged entities are also responsible for deployment (Article 50.3 AIA). They must inform natural persons exposed to biometric categorisation systems of the functioning of the system in the same terms as discussed in relation to emotion recognition systems.

In order for the deployer to be able to comply with its obligations, it must have the necessary information, i.e., be aware that the system is carrying out biometric categorisation. So, in some situations, the responsibility set out in Article 50 should be extended to the system providers. They will know how the system works and can give the required information to the parties that need to report the existence of the classification.

Finally, as in the systems described in the previous section, also in the case of biometric categorisation systems, an exception to the obligation of transparency is provided for when the systems are for the purpose of detecting, preventing and investigating criminal offences, subject to appropriate safeguards for the rights and freedoms of third parties, and in accordance with Union law, where permitted by law.

[51] Edwards, L., *The EU AI Act: a summary of its significance and scope*, Ada Lovelace Institute, 2022.

VIII. Artificial Intelligence systems that generate or manipulate content that constitutes deep fake

1. Covered systems

As we have seen above, Artificial Intelligence makes it possible to generate content (images, videos or voice) or manipulate existing content. Sometimes this content can look very realistic or similar to existing content, which can lead people to believe that it is authentic. The plausibility of the generated content can be so high that, as Seow et al. note, it is necessary to ask whether the aphorism "seeing is believing" is still valid.[52]

The AIA has included among the systems that must comply with transparency obligations those AI systems that generate or manipulate image, audio or video content that constitutes deep fake.

The AIA defines deep fake as AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful.

Ultra-fakes of images or videos may involve creating a non-existent face; transferring the facial expression or body movements of one person to another; manipulating facial attributes (e.g., eye or skin colour) altering a person's appearance; or swapping faces while maintaining the original expressions[53]. These contents have a strong appearance of reality but have never existed or happened.[54]

Image or video manipulation is becoming commonplace as new applications are appearing with more realistic and higher quality results that make it increasingly difficult to distinguish the real from the fake. Also because they are easily accessible to anyone even without technical knowledge and offer surprising results from a single image. In addition, some of these applications are available in open source making them easily accessible.[55]

The spread of deep forgery is being driven by the evolution of Artificial Intelligence but also by the increased availability of databases on which algorithms are trained. All this is resource-intensive.[56]

[52] Seow, J. W., Lim, M. K., Phan, R. C. and Liu, J. K., "*A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities*", *Neurocomputing*, No 513 (2022).

[53] Seow, J. W., Lim, M. K., Phan, R. C. and Liu, J. K., "*A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities*", *op.cit.*

[54] Albahar, M. and Almalki, J., "*Deepfakes: Threats and countermeasures systematic review*", *Journal of Theoretical and Applied Information Technology*, No 97 (2019).

[55] Naitali, A., Ridouani, M., Salahdine, F. and Kaabouch, N., "*Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions*", *Computers*, no 12 (2023).

[56] Seow, J. W., Lim, M. K., Phan, R. C. and Liu, J. K., "*A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities*", *op.cit.*

Deep fakes can be generated or used for a variety of purposes. For example, they are spreading in the multimedia industry (e.g., in the recreation of scenes in films; in the incorporation of special effects; or to dub actors into any language), in video games (e.g., creating virtual doubles of players). It is also being used in education, healthcare, personal assistance or interpreting (e.g., translating a speech and at the same time altering lip movements and facial expressions to simulate that everyone speaks the same language). There are even some applications to help manage grief or to allow interaction with deceased celebrities[57]. They are also finding many applications in business (e.g., for the creation of marketing campaigns, virtual brand ambassadors or models).[58]

However, in recent years, the use of ultra-counterfeits to misinform, defraud or manipulate is multiplying. This problem is significantly increased by the use of social networks[59]. As a result, ultra-counterfeiting is now considered to be one of the greatest threats to society[60], and has led many authorities to promote measures to tackle disinformation.[61]

The use of ultra-counterfeits can serve a wide variety of purposes and can occur in both the public and private spheres. In the public sphere, the misleading use of ultra-falsifications can aim to influence public opinion or electoral results[62], undermine public confidence in institutions[63], widen political polarisation or even support the discourse of extremist groups[64]. This may

[57] Caporusso, N., *Deepfakes for the good: A beneficial application of contentious Artificial Intelligence technology*, Springer, 2021.

[58] Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. and Dwivedi, Y. K., "*Deepfakes: Deceptions, mitigations, and opportunities"*, *Journal of Business Research*, No 154 (2023).

[59] Westerlund, M., "*The emergence of deepfake technology: A review"*, *Technology innovation management review*, no. 9 (2019); Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. and Dwivedi, Y. K., "*Deepfakes: Deceptions, mitigations, and opportunities"*, *op.cit.*; Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A. and Malik, H., "*Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward"*, *Applied Intelligence*, No 53 (2023).

[60] Caldwell, M., Andrews, J. T. A., Tanay, T. and Griffin, L. D., "*AI-enabled future crime"*, *Crime Science*, No 9 (2020).

[61] Examples include the work of the European Union, for example in the Communication Fighting online disinformation: A European approach [COM(2018) 236 final] and, more recently, in various measures included in Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a digital single market for services and amending Directive 2000/31/EC (the Digital Services Regulation).

[62] Naitali, A., Ridouani, M., Salahdine, F. and Kaabouch, N., "*Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions"*, *Computers*, no 12 (2023).

[63] Seow, J. W., Lim, M. K., Phan, R. C. and Liu, J. K., "*A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities"*, op.cit.

[64] Europol, *Facing reality? Law enforcement and the challenge of deepfakes*, 2022.

also affect the credibility of the media, which has the added task of confirming the veracity of the ultra forgeries.

In the private sphere, the generation and dissemination of ultra-counterfeits is also being used, inter alia, to defraud, harass, extort money, take revenge on individuals or to impersonate identities[65]. Some of these actions can cause serious damage to the reputation or credibility of the individuals concerned. They can also create confusion among consumers and have a negative impact on the market (e.g., through the dissemination of ultra-falsified images or videos of company executives in a compromising situation or manipulating statements).[66]

In addition to the errors arising from the deep fakes themselves, another problem linked to their generation is the difficulty of detecting them. AI systems make it possible to generate hyper-realistic deep fakes and to manipulate content by reducing or even suppressing traces that would allow the manipulation to be observed.

A number of measures are being promoted in response to these problems.

Firstly, progress is being made in the development of techniques to assess the authenticity of an image or video or to detect forgeries[67]. In particular, algorithms are being developed to look for inconsistencies between the frames that make up an image (e.g., inconsistencies between speech and lip movement, eye or eyelid movement, leftovers, inconsistencies in the lighting of different parts of the image or reflections of light in the eyes)[68]; or to analyse physical or physiological elements of the image to assess its crdibility (e.g., by analysing the colour of the skin that generates the circulation of blood in the face). However, detection systems are not yet sufficiently reliable, among other aspects, because of the quality and availability of data to train the algorithms[69], because videos as they are disseminated, compressed or reduced are altered[70],

[65] Europol, *Facing reality? Law enforcement and the challenge of deepfakes*, 2022.

[66] Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. and Dwivedi, Y. K., "*Deepfakes: Deceptions, mitigations, and opportunities*", op.cit.

[67] Rana, M. S., Nobi, M. N., Murali, B. and Sung, A. H., "*Deepfake detection: A systematic literature review", IEEE access*, No. 10 (2022).

[68] Westerlund, M., "*The emergence of deepfake technology: A review*", op.cit; Naitali, A., Ridouani, M., Salahdine, F. and Kaabouch, N., "*Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions"*, op.cit.

[69] Naitali, A., Ridouani, M., Salahdine, F. and Kaabouch, N., "*Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions*", op.cit.

[70] Europol, *Facing reality? Law enforcement and the challenge of deepfakes*, 2022.

or because they do not yet work well in real time[71]. The use of technologies, such as blockchain, to verify the legitimacy and origin of content in a trusted, secure and decentralised way is also being promoted.[72]

Secondly, progress is being made in promoting responsible and ethical use of these AI systems to prevent the content generated from contributing to misinformation or the development of criminal or harmful activities[73]. In this direction, greater awareness and training can help to avoid or minimise the harmful effects that may arise from the use of AI systems and AI that generate ultra-falsifications and their dissemination on social networks for the purpose of misinformation.

Thirdly, progress is being made in regulating the use of ultra-counterfeits. To this end, different options have been proposed. One option would have been to limit or ban the circulation of ultra-counterfeits. However, while this solution can prevent or avoid certain damage, it can also generate new impacts[74]. Indeed, we cannot ignore the fact that the use of these AI systems can be a manifestation of freedom of expression or freedom of artistic creation. In the face of these options, and regardless of what may be derived from the consideration of certain systems as having a high impact, the AIA has opted to provide for compliance with the transparency obligations that are analysed in the next section.

2. Scope of obligations

Article 50.4 AIA provides for the obligation to make public that content or images have been artificially generated or manipulated.

The regulated entities are those responsible for the deployment of the AI system that generates or manipulates images, audio or video content that constitutes forgery. As noted above, in certain circumstances, deployers may be unaware of the extent to which a particular piece of content has been created by an AI system. Given this lack of knowledge, it may be difficult for them to effectively provide the information to the final recipients of the image or video.

[71] Naitali, A., Ridouani, M., Salahdine, F. and Kaabouch, N., "*Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions*", op.cit.

[72] Rana, M. S., Nobi, M. N., Murali, B. and Sung, A. H., "*Deepfake detection: A systematic literature review", op.cit.*

[73] Naitali, A., Ridouani, M., Salahdine, F. and Kaabouch, N., "*Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions", op.cit.*

[74] Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A. and Dwivedi, Y. K., "*Deepfakes: Deceptions, mitigations, and opportunities", op.cit.*

The transparency obligation is for deployers to disclose that the content has been artificially generated or manipulated.

The AIA seeks to strike a balance between freedom of expression and the prevention of disinformation and manipulation through the generation or dissemination of deep fakes. To this end, it provides that "Where the content forms part of a manifestly creative, satirical, artistic or fictional work or programme, the transparency obligations set out in this paragraph shall be limited to the obligation to make public the existence of such artificially generated or manipulated content in an appropriate manner that does not impair the exhibition or enjoyment of the work".

The AIA also seeks to ensure freedom of the press. To this end, it provides that "Those responsible for the deployment of an AI system that generates or manipulates text that is published for the purpose of informing the public on matters of public interest shall disclose that the text has been artificially generated or manipulated".

As in other cases, exceptions to this obligation are those cases authorised by law to detect, prevent, investigate or prosecute criminal offences. Also, when the content generated has been reviewed by a person or subject to editorial control and when a natural or legal person has editorial responsibility for the publication of the content.

IX. Recapitulation

Article 50 AIA regulates transparency obligations aimed at preventing certain AI systems, which in themselves must not pose a risk to the interests of the European Union or to the fundamental rights of individuals, from being used in such a way that they do not cause confusion, deception, manipulation, or misinformation among persons interacting with such systems or recipients of the content generated or manipulated.

To this end, it foresees that providers or those responsible for deployment, as the case may be, must provide information to users or affected persons so that they can be aware of or avoid a given situation or the harmful results that may eventually be caused by the use of the AI system.

In practice, the transparency obligations under Article 50 AIA, aimed at ensuring communication to the user or affected person of the existence of an AI system or of an artificially generated result, will in many cases be complementary to other transparency obligations under the AIA, e.g., for high-risk AI systems, which aim to ensure traceability of the functioning of the AI system and the explainability of the results obtained. In order to achieve the intended purpose, it is essential that the information provided is sufficiently clear for the intended purpose to be achieved.

The Commission's periodic assessment under Article 112.2 of the AIA will assess whether the obligations under Article 50 have been achieved and will determine whether there is a need to amend the list of AI systems that must comply with the transparency obligations.

Sandbox, governance, oversight, sanctions, rights and confidentiality in the Regulation

SANDBOX, CONTROLLED SPACES AND REAL-WORLD TESTING OF ARTIFICIAL INTELLIGENCE SYSTEMS IN THE REGULATION. MEASURES FOR SMES, STARTUPS AND MICRO-ENTERPRISES

Lorenzo Cotino Hueso

Professor of Constitutional Law at the University of Valencia. Valgrai[1]

I. The "Measures in support of innovation" of Chapter VI

Chapter VI, under the title "Measures in support of innovation", includes seven articles of considerable length, totaling more than 5,500 words, addressing various topics. Essentially, since the Commission's initial proposal in 2021, this chapter regulates "Controlled AI Test Spaces", which will be abbreviated here as "sandboxes". In addition to regulating their existence and regime (Art. 57-58), the possible processing of personal data by AI systems in these sandboxes is legitimised (Art. 59). Although not in the Commission's initial proposal, the AI Act includes the regulation of "Testing of high-risk AI systems in real world conditions outside AI regulatory sandboxes" (Art. 60), together with an article on the "informed consent" of individuals affected by such testing (Art. 61). From the outset, Chapter VI also contained an article on "Measures for providers and deployers, in particular SMEs, including start-ups" (Art. 62), to which an article on "Derogations for specific operators" has been added (Art. 63).

Among the most notable changes since the initial version, apart from the inclusion of real-world testing, one of the most noteworthy modifications from the original draft is the requirement to create a sandbox in every state within two years of the AI Act's implementation.AI Act. It also provides for the possibility of creation at regional or local level, jointly with other states, as well as by the Commission and the European Data Protection Supervisor. Likewise, the Council also introduced the aims or objectives of sandboxes

[1] cotino@uv.es. OdiseIA. This study is the result of research from the following projects: MICINN Project "Public rights and guarantees against automated decisions and algorithmic bias and discrimination" 2023-2025 (PID2022-136439OB-I00) funded by MCIN/AEI/10.13039/501100011033/; "The regulation of digital transformation…" Generalitat Valenciana "Algorithmic law" (Prometeo/2021/009, 2021-24); "Algorithmic Decisions and the Law: Opening the Black Box" (TED2021-131472A-I00) and "Digital transition of public Administrations and Artificial Intelligence" (TED2021-132191B-I00) of the Recovery, Transformation and Resilience Plan. CIAEST/2022/1, Research Group in Public Law and ICT, Catholic University of Colombia; MICINN; CIAEST/2022/1, Digital Rights Agreement-SEDIA Scope 5 (2023/C046/00228673) and Scope 6 (2023/C046/00229475).

(Art. 57.9). The regime of flexibility in the compliance of sandbox participants and their possible liability has also been changing. Elements such as the specific plan to be submitted, the consequences of participation in a sandbox, the obligation to provide written proof of the activities carried out by the participant, and the exit report have been added. It is relevant that participation in a sandbox has been linked as a means to demonstrate compliance with the conformity assessment process.

On the other hand, in this Chapter VI, the Parliament (Amendment 516) proposed to include an article on "Promotion of AI research and development in support of socially and environmentally beneficial outcomes", with promotion mandates. However, this proposal did not succeed.[2]

II. Origin and concept of sandboxes and controlled spaces

Montesquieu stated that "sometimes it is even convenient to test a law before establishing it. The constitutions of Rome and Athens were very wise in this respect: the decisions of the Senate had the force of law for one year, and only became perpetual by the will of the people"[3]. The United States allowed the "states-as-laboratories". As Justice Brandeis stated, "One of the happy incidents of our federal system is that a single brave State may, if its citizens so choose, serve as a laboratory and try new social and economic experiments without risk to the rest of the country" (Justice Brandeis' dissenting opinion in New State Ice v. Liebmann 285 U.S. 262, 310 [1932]).[4]

Despite various historical experiences in testing regulations and innovation for centuries, it wasn't until the second half of the 20th century that they developed, accompanying the interventionism of the social state in social and economic life.[5] However, the first formal regulatory sandbox and the spread of the

[2] It included a mandate to promote AI solutions that improve accessibility for people with disabilities, reduce socio-economic inequalities and support sustainability and environmental protection objectives. This included measures such as providing priority access; allocating public funding to AI projects with positive social and environmental impact; organising AI Act awareness-raising activities; specific funding and application procedures, adapted to the needs and specific accessible communication channels. Civil society participation was also encouraged with regard to AI for society and the environment.

[3] Doménech Pascual, G., "Las regulaciones experimentales", *Anuario del buen gobierno y de la calidad de la regulación,* (monograph on sandbox Ponce Solé, J. and Villoria Mendieta, M., coords.) n.º 1, 2022, pp. 103-146. Cites Montesquieu, C.-L. de S. *Del espíritu de las leyes* translation by Blázquez y de Vega, Tecnos, Madrid 2000.

[4] Separate opinion of Justice Brandeis in New *State Ice v. Liebmann 285 U.S. 262,310* (1932). *Ibid.*

[5] BMWi, *Making Space for Innovation: The Handbook for Regulatory Sandboxes*, German Feder-

concept globally occurred to test the market introduction of *Fintech* products.[6] Since 2014, by the UK's *Financial Conduct Authority* (FCA), it has been extended to other regulated sectors, such as healthcare (supervised by the Care Quality Commission), energy (OfGem), and from there to other sectors.[7] The success seems to be proven since, in July 2023, the OECD stated that there have already been a hundred sandbox initiatives, including in *fintech* and privacy.[8]

The terminology used to refer to sandbox-like realities or controlled test spaces is very varied: 'living laboratories', 'innovation spaces', 'regulatory test beds' or 'real life experiments'. Years ago, in its definition, the Council of the EU stressed that '8. perceives regulatory sandboxes as concrete frameworks which, by providing a structured context for experimentation, enable, where appropriate, in a real-world environment, the testing of innovative technologies, products, services or approaches – at the moment especially in the context of digitalisation – for a limited time and in a limited part of a sector or area under regulatory supervision ensuring that appropiate safeguards are in place'. Also, "9. understands *experimental clauses* as legal provisions which enable the authorities tasked with implementing and enforcing the legislation to exercise on a case-by-case basis a degree of flexibility in relation to testing innovative technologies, products, services or approaches".[9]

In Germany, "regulatory "sandboxes" are test areas established for a limited time, covering a limited area, in which innovative technologies and business models can be tried out in real life".[10] And the experimentation clauses as a technical regulatory instrument that allows for exceptions to the general

al Ministry for Economic Affairs and Energy, 2019, p. 7 https://www.bmwk.de/Redaktion/EN/Publikationen/Digitale-Welt/handbook-regulatory-sandboxes.pdf?__blob=publicationFile&v=1

[6] On Spanish regulation in this area, Huergo Lora, A. "Un "espacio controlado de pruebas" (regulatory sandbox) para las empresas financieras tecnológicamente innovadoras". El "Anteproyecto de Ley de Medidas para la transformación digital del sistema financiero", in *El Cronista del Estado Social y Democrático de Derecho*, n.º 76 (September), 2018, pp. 48-59 and Hernández Peña, J. C., "La propuesta de un sandbox regulatorio para el sector financiero español: ¿más luces que sombras?", *Revista General de Derecho de los Sectores Regulados* n.º 2, 2018.

[7] In this regard, Truby, J., "Decarbonizing Bitcoin: Law and policy choices for reducing the energy consumption of Blockchain technologies and digital currencies", *Energy Research & Social Science*, Volume 44, 2018, pp. 399-410, https://doi.org/10.1016/j.erss.2018.06.009.

[8] OECD, *Regulatory sandboxes in Artificial Intelligence*, OECD Digital Economy Papers, July 2023 No. 356, 2023, p. 8 https://read.oecd.org/10.1787/8f80a0e6-en?format=pdf

[9] Council of the European Union, *Council Conclusions on Regulatory sandboxes and experimentation clauses as tools for an innovation-friendly, future-proof and resilient regulatory framework that masters disruptive challenges in the digital age*, Brussels, 16 November 2020 (OR. en), 13026/20https://data.consilium.europa.eu/doc/document/ST-13026-2020-INIT/en/pdf

[10] BMWi, *Making Space for* Innovation… cit. p. 7.

legal framework. They therefore allow new approaches to be adopted, without being able to predict the outcome. And they offer the opportunity to learn about laws and their effects.[11]

For its part, the OECD notes that "AI regulatory sandboxes should be seen as one of several tools for regulatory experimentation and innovation, along with complementary areas: standardisation, innovation centers, other controlled-test spaces such as those for financial technologies and privacy, and governance technologies".[12]

Article 3(55)(1) of the AI Act defines a "AIregulatory sandbox" as "a controlled framework set up by a competent authority which offers providers or prospective providers of AI systems the possibility to develop, train, validate and test, where appropriate in real-world conditions, an innovative AI system, pursuant to a sandbox plan for a limited time under regulatory supervision".

III. Artificial Intelligence sandbox experiences

The experience of AI-linked sandboxes is clearly associated with compliance with data protection regulations, especially in Europe (UK, Norway and France), as well as in Colombia. In other contexts, sandboxes have been particularly linked to regulatory experiments in the *Fintech sector*.

The UK ICO launched a sandbox in 2019 "to support organisations developing particularly innovative products or services that process personal data"[13]. In the context of AI, the focus was on "Exceptional Innovations", "emerging technologies" (such as consumer health technology, wearable devices and software applications that help people assess their health and well-being; Internet of Things (IoT), immersive technology; decentralised finance: software that uses *blockchain* technology to support peer-to-peer financial transactions) and on "Biometrics". You can follow the exit reports of all participants since then. In November 2021 they published a Beta Report on the learnings from the Sandbox.[14] As discussed below, it is notable that no specific exceptionalities or particularities for participants were regulated.

[11] *Ibid*, p. 81.

[12] OECD, *Regulatory sandboxes… cit.* pp. 24 ff.

[13] https://ico.org.uk/for-organisations/advice-and-services/regulatory-sandbox/

Particularly worth following is the ICO, Information Commissioner's Office, *The Guide to the Sandbox*, https://ico.org.uk/for-organisations/regulatory-sandbox/the-guide-to-the-sandbox.

[14] https://ico.org.uk/about-the-ico/news-and-events/news-and-blogs/2019/03/ico-opens-sandbox-beta-phase-to-enhance-data-protection-and-support-innovation/

In Norway, the Norwegian *Data* Protection Authority, *Datatilsynet*, created in 2021 a "Sandbox for Artificial Intelligence"[15] inspired by the one in the UK. Nor does this sandbox regulate a special regime for participants, as can be seen below. Participants must follow the Ethical Guidelines for Responsible AI of the EU's High Level AI Expert Group.[16] It received 25 applications from multiple public and private organisations and selected four projects for the sandbox, which started in March 2021. Results reports were disseminated in 2023.[17]

France has extensive experience and general and even constitutional regulation of sandboxes.[18] In AI, the data protection authority (CNIL) has been a clear leader. In 2021, the sandbox was dedicated to health applications, with 10 projects.[19] It did not exempt compliance with the GDPR, but the sandboxaimed to facilitate such compliance. The 2022 edition was dedicated to education technology with four projects.[20] In July 2023, the sandbox focused

[15] https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/

[16] HLEG-European Commission, *Ethical Guidelines for Reliable AI*, 2019, *Ethical Guidelines for Reliable AI*, 2019, https://op.europa.eu/es/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

[17] https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/reports/

On transparency of data protection in AI systems. The importance of setting the purposes of the systems' data processing is pointed out (Ruter Report). The *Ahus* Report especially analyses the algorithmic discrimination of an algorithm for predicting heart failure. The *Simplifai* Report focuses on the uses of AI administration for recording and archiving emails or decision making (NVE). The *Finterai* Report focuses on federated and limited access to data for learning in the fight against money laundering and terrorist financing. The *AVT* Report on individual assessments and tailored education with privacy. The *NAV* Report, an AI tool for predicting the development of sick leave at the individual level, is also worth considering.

[18] For all, Conseil d'État, *Les expérimentations: comment innover dans la conduite des politiques publiques*, Conseil d'État, Paris, 2019, https://www.conseil-etat.fr/Media/actualites/documents/2019/10-octobre/etude-pm_experimentations_vdef.

[19] Results at https://www.cnil.fr/en/digital-health-and-edtech-cnil-publishes-results-its-first-sandboxes#:~:text=its%20first%20%E2%80%9Csandboxes%E2%80%9D-,Digital%20health%20and%20EdTech%3A%20the%20CNIL%20publishes,results%20of%20its%20first%20%E2%80%9Csandboxes%E2%80%9D&text=The%20CNIL%20publishes%20the%20recommendations,health%20and%20educational%20digital%20tools

[20] https://www.cnil.fr/en/edtech-sandbox-cnil-supports-10-innovative-projects Projects Daylindo, Klassroom, France Université Numérique and "personal cloud" Academy of Rennes. Report at https://www.cnil.fr/sites/cnil/files/2023-07/bilan_bac_a_sable_edtech.pdf

on three projects related to Artificial Intelligence in public services.[21] In 2023, the CNIL launched an action plan on AI.[22]

Other countries in the EU have not focused on data protection. Germany's AI strategy included living labs and AI test beds, creating new experimentation clauses as a legal basis. In the field of automated driving, some tests have been developed, such as the sandbox project in North Rhine-Westphalia[23]. A digital country like Estonia launched in 2022 a test bed (*AI Govstack Testbed*)[24] focusing on data development, management, analysis and labeling. The Malta *Digital Innovation Authority* (*Malta Digital Innovation Authority*) created in 2020 an *MDIA-TAS* (*Technology Assurance Sandbox*), a regulatory sandbox focused on emerging technologies such as AI.[25] This sandbox aims to help companies comply with existing regulations.

In Switzerland, the Canton of Zurich developed the *Innovation Sandbox for Artificial Intelligence*[26] to assist with regulatory issues and enable the use of new data sources. It does not appear to have specific regulatory coverage. Unlike other sandboxes, "selected projects are not only being reviewed, but also implemented". Between March and June 2022, 21 AI projects were submitted, of which five were selected and are currently in the implementation phase. It will last until April 2024 and there will be another call between March and May 2024. Five projects have been selected:[27] autonomous systems, such as self-driving tractors or lawnmowers in public spaces; infrastructure maintenance with drones; AI applications in education; smart parking in cities; and best practices for privacy by design and machine translations for public administration.

In January 2024*, the AI Sandbox Summit* was held *in Zurich*,[28], bringing together initiatives from Germany, Belgium, Norway, the UK, France and Spain. The summit underlined the importance of regulatory sandboxes and did not consider terminological dispersion or consensus on definitions to be crucial, but rather the adoption of different types according to the needs of

[21] https://www.cnil.fr/en/sandbox-cnil-launches-call-projects-artificial-intelligence-public-services

[22] https://www.cnil.fr/en/artificial-intelligence-action-plan-cnil

[23] The federal state of North Rhine-Westphalia (NRW), where an extensive Digi-Sandbox.NRW project is under development. Its website lists several reallabs in NRW, but none focuses on privacy protection and Artificial Intelligence.

[24] https://e-estonia.com/ai-govstack-testbed_eng/

[25] https://www.mdia.gov.mt/technology-assurance-sandbox/

[26] https://www.zh.ch/en/wirtschaft-arbeit/wirtschaftsstandort/innovation-sandbox.html https://innovation.zuerich/en/sandbox/

[27] The dossiers can be accessed on the aforementioned website.

[28] https://www.zh.ch/en/wirtschaft-arbeit/wirtschaftsstandort/innovation-sandbox.html

each country. International collaboration was encouraged and the creation of a common database of relevant use cases in the different European sandboxes is foreseen to facilitate the exchange of knowledge.

In 2021, under Federal Law No. 258-FZ, Russia introduced regulatory sandboxes to encourage digital innovation. Among the eight selected projects, there were AI applications in the fields of transportation, healthcare and tourism.[29]

In Ibero-America, in 2021, the government of Chile published a document on AI sandboxes.[30] The Government of Colombia created a guide on AI regulatory sandboxes in 2020[31], and the Colombian Authority for the Protection of Personal Data launched a regulatory sandbox on privacy by design and by default in Artificial Intelligence projects.[32] Two projects were selected in 2021 (*NaaS* Colombia S.A.S., - "Evolución Index Core" and Alcaldía de Barranquilla - "*Chatbot*") and in 2022 the "*Diyosoy*" of *Wolman Group de Colombia Limitada*. Granero recalls that in Argentina, for the City of Buenos Aires, Law 6491 on Controlled Test Space of 9 December 2021 regulated the framework for this type of testing.[33] In the province of Mendoza, Law No. 9086 of 30 July 2018,[34] in its articles 52 and following, on urban transport through mobile applications and platforms, was considered by the Supreme Court of Justice of Mendoza as an experimental regulation.[35]

In Asia, the *Monetary Authority of Singapore* launched its Fintech regulatory sandbox in 2018, facilitating the testing of AI applications.[36] The principles published by this authority were incorporated into its National AI Strategy. On 1 April 2019 in Korea, the Special Act on Assistance to Financial Innovation (*SAAFI*) came into force to support the development of financial services and increase benefits for consumers. In addition to a financial sand-

[29] https://a-ai.ru/en

[30] Guño, A., *Artificial Intelligence Regulatory Sandbox in Chile. Discussion paper.* CAF, Development Bank of Latin America, August, 2021, https://www.economia.gob.cl/wp-content/uploads/2021/09/PaperSandboxIA.pdf

[31] Superintendencia de Industria y Comercio, *Sandbox on privacy by design and by default in Artificial Intelligence projects*, SIC, Colombia, 2021, https://www.sic.gov.co/content/sandbox-sobre-privacidad-desde-el-disen%CC%83o-y-por-defecto-en-proyectos-de-inteligencia-artificial.

[32] https://www.sic.gov.co/sandbox-microsite (not available).

[33] https://documentosboletinoficial.buenosaires.gob.ar/publico/ck_PL-LEY-LCABA-LCBA-6491-22-6295.pdf

[34] https://www.mendoza.gov.ar/gobierno/wp-content/uploads/sites/19/2018/10/Ley-de-Movilidad-N%C2%BA-9086.pdf

[35] Granero Horacio R., "La imperiosa necesidad de regular -bien- la inteligencia artificial", paper, FACA, *ElDial*, 2023.

[36] https://www.mas.gov.sg/development/fintech/regulatory-sandbox

box,[37] various ministries in Korea created in 2019 a public industrial experimentation framework with a limited regulatory exemption for companies to test innovative products, services and business models. There were seven AI+X projects.

At the global level, the Global *Financial Innovation Network* (GFIN) financial sandbox started in 2020 with more than 50 financial institutions from all over the world and some lessons learned.[38] However, joining has proven challenging due to the complex compatibility between legal regimes. Thus, out of 38 applications from companies, only two were successful (*Banksysteme and Bedrock AI*).

IV. The advantages of Artificial Intelligence sandboxes from different points of view

Article 57.5 notes the general objective of an AI sandbox: "provide for a controlled environment that fosters innovation and facilitates the development, training, testing and validation of innovative AI systems for a limited time before their introduction to the market or their entry into service". Furthermore, Article 57. 9th AI Act points out the possible objectives of a sandbox: to foster innovation and competitiveness, to facilitate the development of an AI ecosystem, to "facilitate and accelerate market access", in particular for SMEs and startups, to "improve legal certainty and contribute to the sharing of best practices through cooperation with authorities". In addition, it mentions "sharing best practices through cooperation with the authorities involved" and "improving legal certainty to achieve compliance with this Regulation". The Council version also stated "contributing to the uniform and effective application of the IAM" and "contributing to the development or updating of harmonised standards". In its versions prior to the one adopted, Article 53(1)(a) AI Act stated that "The AI regulatory security enclosure shall allow and facilitate the participation of notified bodies, standardization bodies and other relevant stakeholders where appropriate".

According to the OECD,[39] for the case of AI, it is understood that sandboxes are particularly suitable for testing the readiness of AI-related products or services for commercialization in the light of standards and norms. The OECD underscores the interdependence between standards and AI policy

[37] https://sandbox.fintech.or.kr/?lang=en

[38] The reports can be accessed at https://www.thegfin.com/crossborder-testing

[39] OECD, *Regulatory sandboxes… cit.* p. 24.

regulation, particularly in the precise risk-based approaches that standards employ to enforce regulation. Thus, we can use data collected in a sandbox to detect patterns and identify the need for a standard in a specific area. On the other hand, AI sandboxes can use standardization processes for testing.

Multi-disciplinary and multi-stakeholder cooperation is essential for an AI sandbox due to the strong cross-cutting nature of AI, which has concurred, in particular, with financial technologies and the data protection domain. This cooperation requires avoiding "silos"[40] and encouraging participation between the different regulatory authorities involved: "competition authorities, intellectual property offices, national standardisation bodies and data protection authorities". In this respect, some examples of cooperation in Korea, Germany and Brazil are mentioned.[41] Coordination with market players, including business is also important.

Sandboxes, and AI in particular, offer a variety of advantages from different perspectives From public interests and purposes, there are advantages in terms of innovation, for authorities and the improvement and implementation of regulation. From the economic and market point of view, there are also various advantages. While there are numerous public or collective interests, the interests of the entities participating in an AI sandbox are also varied, especially if they are small and medium-sized enterprises and startups.

From an innovation point of view, an AI sandbox involves the promotion of innovation through the provision of data,[42] the transfer of *know-how* and enablement of new projects, and shared governance in collaboration of the private sector with science, academia and the regulatory sector. The sandbox model fosters research-policy partnerships, consensus building and policy solutions. In a sandbox, multidisciplinary approaches take place in controlled contexts and under the authority of a professionalised bureaucracy, reducing information asymmetries. The OECD has stressed that sandboxes should not only be used to validate the expectations of binding legislation, but also to support innovation and innovation hubs. In this line, the AI Act points to

[40] *Ibid*, pp. 12-13 and especially p. 19. Coordination of regulatory responses between national agencies is critical, Brummer, C. and Y. Yadav (2019), "Fintech and the Innovation Trilemma", *The Georgetown Law Journal* 107, https://scholarship.law.vanderbilt.edu/faculty-publications/1084/.

[41] *Ibid*, pp. 19 ff. Korea's sandbox has inter-ministerial participation, and various sectors in Germany use flexible and generic sandbox frameworks. In Brazil, the Securities and Exchange Commission and the Central Bank created an internal committee that interacts with universities, researchers, associations and industry representatives to evaluate sandbox applications.

[42] Followed by Canton of Zurich, *Innovation Sandbox for Artificial Intelligence (AI),* https://www.zh.ch/en/wirtschaft-arbeit/wirtschaftsstandort/innovation-sandbox.html

technical and scientific support from innovation hubs in the AI ecosystem (Consid. 139).

For authorities and regulatory reform, the practice of sandboxing increases the speed of business approvals,[43] improves communication between regulators and businesses,[44] enables assessment of the effectiveness of regulation and policy under real-world conditions, and generates empirical data that can be used for better regulatory decision-making or other public interest purposes. A sandbox can enable more adaptive and dynamic regulation.[45] As noted in Switzerland, they can "provide regulatory clarity". Thus, a sandbox makes it possible to learn how to implement regulation by authorities, how to regulate new types of services in collaboration with private actors, and to gather more information. It allows suggesting or recommending regulatory adjustments, setting criteria to facilitate compliance with regulations, and outlining procedures for enforcement. It also makes it easier to make regulatory compliance an essential component of the design and implementation of Artificial Intelligence projects.[46]

From a market and economic perspective, sandboxes showcase the sandboxing country's innovative economy. They also enhance competitiveness in the AI landscape to develop new applications, encourage experimentation, or even develop technologies that can be used outside the country. Sandboxes facilitate the testing of new products that would otherwise not have access to

[43] *Ibid.* Thus, it is noted that since its launch in 2015, the UK FCA's sandbox has supported more than 700 firms and increased their average speed to market by 40% compared to the regulator's standard authorisation time. Reference is made to Truby, J. et al., "A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications", *European Journal of Risk Regulation*, 2021 https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/sandbox- approach-to-regulating-highrisk-artificial-intelligence-applications/C350EADFB379465E7F4A95B973A4977D.

[44] Among other sources, see Ranchordás, S., "Experimental Regulations for AI: Sandboxes for Morals and Mores", *Morals & Machines*, 1(1), pp. 86-100. https://doi.org/10.5771/2747-2021-1-86 and Superintendencia de Industria y Comercio, *Sandbox sobre privacidad… cit.*

https://www.sic.gov.co/sites/default/files/normatividad/112020/031120_Sandbox-sobre-privacidad-desde-el-diseno-y-por-defecto.pdf

[45] Ranchordás, S., "Experimental Regulations for AI…" *cit.*; Guño, A., *Sandbox Regulatorio… cit.* p. 19 and Superintendencia de Industria y Comercio, *Sandbox sobre privacidad… cit.*).

[46] Thus, it is noted that one of the UK FCA's fifth fintech sandbox cohort projects resulted in regulatory changes. Specifically, progress in this regard is noted in the *UK FCA Policy Statement PS19/22* Guidance on cryptoassets. Also adaptation to anti-money laundering and counter-terrorist financing control measures for remote customer onboarding initiatives by the Hong Kong Monetary Authority 2020. Also, Joint Consultation Paper 2019 21-402 of the Canadian Securities Administrators and the Investment Industry Regulatory Organisation of Canada: Proposed Framework for Cryptoasset Trading Platforms.

markets. Fintech and the UK have demonstrated that testing attracts investment.[47] The availability of authorities to learn about new technologies is in itself an element of investment attraction.

From the point of view of the participating entities, the sandbox provides a beneficial legal regime for a limited period of time. In addition, the advantages of participating in these experimentation contexts have been pointed out by the ICO[48] and other entities: access to experience, facilitating learning by the parties, greater confidence in the compliance of their finished product or service, a better understanding of future AI Act regulation and how it affects their company or entity, being held accountable and proactive in their approach to future AI Act regulation by customers, other organisations and responsible regulatory authorities, leading to greater consumer confidence in their organisation and contributing to the development of products and services that can demonstrate their value to the public. The OECD[49] also reports that 40% of enterprises who completed the UK FCA's inaugural financial sandbox program, subsequently secured finance, with investment in financial technology being 6.6 times greater. Furthermore, participation in a sandbox fosters beneficial internal dynamics inside the member organisations.

In the case of small and medium-sized enterprises and startups, there are private and public interests at stake. The cost of liability for damages and impacts of technology, along with the uncertainties of the applicable legal framework, can stifle innovation, a risk that small and medium-sized enterprises and startups cannot afford. Relaxation of strict liability in the context of a sandbox may be relevant.[50] SMEs likely have higher total compliance costs than large firms due to economies of scale, accounting for 17% of total AI investment costs.[51] SMEs also face specific entry barriers such as standards or quality procedures. This is why the EU Council "stresses that regulatory "sandbox-

[47] In the UK, 30% of the venture firms that participated in the regulatory sandbox received venture investment, and the average amount of investment increased 6.6 times. Followed by Guño, A., *Regulatory Sandbox… cit.*

[48] ICO, *The Guide to the Sandbox…* cit. Also, Truby, J. et al. "A Sandbox Approach…" *cit.*

[49] Reference is made to studies UK FCA, *Regulatory sandbox lessons learned report*, 2017, p. 22, https://www.fca.org.uk/publication/research-and-data/regulatory-sandbox-lessons-learned-report.pdf and Goo, J. and J. Heo, "The Impact of the Regulatory Sandbox on the Fintech Industry, with a Discussion on the Relation between Regulatory Sandboxes and Open Innovation", *6 J. Open Innov. Technol. Mark. Complex*, 2020p. 19 https://www.mdpi.com/2199-8531/6/2/

[50] In this regard Truby, J. et al. "A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications", *European Journal of Risk Regulation.* 2022; 13(2), pp. 270-294. doi:10.1017/err.2021.52

[51] Followed by Truby, J. et al. "A Sandbox Approach…" *cit.* note 59, 155, 160, 166.

es" can offer important opportunities, in particular to innovate and grow, for all enterprises, especially SMEs, including micro-enterprises and start-ups, in industry, services and other sectors".[52]

In this direction, the AI Act affirms free access to sandboxes (Art. 58.2 d),[53] and Article 62 specifies that "priority access to controlled AI testing areas shall be given, provided that they meet the eligibility conditions and selection criteria". Furthermore, Article 58(3) recalls that the Commission implementing acts on sandboxes' shall offer potential providers participating in controlled AI testing spaces, in particular SMEs and start-ups, where appropriate, pre-deployment services, such as guidance on the implementation of this Regulation, other value-added services such as assistance with standardisation documents and certification, and access to testing and experimentation facilities, European Digital Innovation Centres and centres of excellence'.

V. The Regulatory Framework of an Artificial Intelligence Sandbox under the Regulation

The AI Act aims to avoid fragmentation in AI regulation among Member States, especially in the area of sandboxes. It had been questioned whether there could be dispersed national regulation: "how a national regulator can fully participate in a regulatory sandbox when the area of regulation falls partly or wholly under EU competences".[54] Among other things, there was a need to avoid AI developers choosing EU states with less stringent sandbox regimes, as well as dispersion in experimental data collection methods and limits. Therefore, a common EU regulation combined with state regulation was advocated.[55] However, it is noted that the proposed regulation could create confusion in the market, as it allows Member States' competent authorities to agree on a common implementation and framework combining EU's and Member States' rules for a sandbox.

Thus, Article 53. 1st AI Act of the Commission proposal mentioned "compliance with the requirements laid down in this Regulation and, where applicable, in other Union and Member State legislation supervised in the

[52] Council of the European Union, *Council Conclusions…* cit.

[53] "without prejudice to any exceptional costs that the competent national authorities may recover in a fair and proportionate manner;".

[54] Yordanova, K., "The shifting sands of regulatory sandboxes for AI" (KU Leuven, Centre for IT&IP Law, 2019) https://www.law.kuleuven.be/citip/blog/the-shifting-sands-of-regulatory-sandboxes-for-ai/ I follow by Truby, J. et al. "A Sandbox Approach… *cit.*

[55] *Ibid.*

framework of the controlled testing area". However, this reference has disappeared, and the trend towards homogeneity in the regulation of AI sandboxes is clear. For example, Article 57 with its 17 paragraphs implies a common regime applicable to AI sandboxes in the EU. In particular, Article 58 (1) has been introduced: "In order to avoid fragmentation across the Union, the Commission shall adopt implementing acts containing common principles and ensuring a range of elements". These common elements to be ensured by the European Commission's acts are detailed in thirteen paragraphs. It is stated that the management of AI sandboxes should ensure "flexibility to establish and manage their controlled AI test spaces" (Art. 58. 2º c).

Therefore, in the EU, any AI sandbox must adhere to the AI Act's regulation of sandboxes AI Actand any future Commission implementing acts (Art. 58. 1). Existing national regulation may not be contrary to the AI Act. However, Article 57(4) states that "This Article shall not affect other regulatory sandboxes established under Union or national law". This implies that other sandboxes whose essential purpose or object is not AI will be governed by general national sandbox rules and, where appropriate, by their specific rules. In any case, it is advised that "where appropriate, the relevant competent authorities in charge of such other controlled sandboxes should consider the benefits of using them also for the purpose of ensuring compliance of AI systems with this Regulation" (Consid. 139). In these cases, it is understood that sandboxes would fall under the AI Act regime.

It is important to consider *the establishment and regulatory framework of an AI sandbox*. In Spain, there has been no general regulation of sandboxes or controlled test spaces until Article 16 of Law 28/2022, of 21 December, on the promotion of the start-up ecosystem. While this general regulation specifically targets start-ups, it serves as a general framework for sandboxes in Spain. In the municipal sphere, the pioneering Municipal Ordinance regulating the Urban Sandbox of the City of Valencia in April 2024 stands out.[56] There are scattered sectoral laws on testing, pilots and sandboxes in the fields of telecommunications[57], the financial sector,[58], the energy sector,[59] for the pub-

[56] The preliminary draft available at https://sede.valencia.es/sede/descarga/doc/DOCUMENT_1_20230005159164. https://sede.valencia.es/sede/ordenanzas/detalle/MzE2NjQ.AvPAlt3D.AvOvTok

[57] Thus, art. 61.f) of Law 9/2014, of 9 May, General Telecommunications regulated authorisations to use the public radioelectric domain for experimental purposes, now in article 86 f) Law 11/2022, of 28 June, General Telecommunications.

[58] In any case, the financial sector was a pioneer and articles 4 to 18 of Law 7/2020 of 13 November for the digital transformation of the financial system should be taken into account.

[59] In the energy field, the 23rd additional provision of Law 24/2013, of 26 December, on

lic sector in the law on science, technology and innovation,[60], evaluation of public policies[61] or on agreements with entities for pilot testing in Catalonia.[62]

As mentioned above, the final version of the AI Act obliges each State to implement at least one national-level AI sandbox within 24 months of its entry into force (art. 57.1). The establishment of a specific AI sandbox in Spain must be carried out by regulation,[63] on the basis of the recent general legal coverage. In the field of AI in Spain, there exists a Royal Decree 817/2023 of 8 November,[64] which theoretically established an AI sandbox. However, it seems to be a failed or abandoned instrument. Once the AI Act is adopted, the AI sandbox in Spain could only be reactivated with a reform of Royal Decree 817/2023, which would have to comply with the AI Act regulation. Otherwise, it would also not be valid as the mandatory sandbox needs to be established in the first 24 months.

The regulation establishing the AI sandbox in Spain will set the particular regulatory framework for the sandbox. In addition to said regulations, it will not be unusual for rules or general conditions for the call to be issued. In general, the terms and conditions are a general administrative act with a defined duration.[65] This instrument is suitable for establishing the rules of the sandbox which, at the same time, initiate the process of access and selection of participants. Another formula for the concrete regulation of a sandbox is the

the Electricity Sector (according to Royal Decree-Law 23/2020, of 23 June) and Royal Decree 568/2022, of 11 July, which establishes the general framework of the regulatory test bed for the promotion of research and innovation in the electricity sector.

[60] Law 14/2011 of 1 June 2011 on Science, Technology and Innovation, as amended by Law 17/2022 of 5 September 2011, should also be taken into account. Article 33.1 of this law provides for innovative measures through accelerators, incubators and demonstration centres; experimentation and dissemination spaces; public procurement of innovation; and framework service agreements for the development of solutions involving the introduction of disruptive technologies in the administration (Art. 33.1.k).

[61] It is also worth taking into account Law 27/2022 of 20 December on the institutionalisation of public policy evaluation in the General State Administration.

[62] In the case of Catalonia, article 64.4 of Law 19/2014, of 29 December, on transparency, which refers to agreements with entities for pilot tests, must be taken into account.

[63] Art. 16.1º Law 28/2022 of 21 December: "The public authorities shall promote, by regulation, the creation of controlled environments".

[64] "establishing a controlled test environment for testing compliance with the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence".

[65] In the context of calls for public personnel, they are administrative acts (Royal Legislative Decree 5/2015, of 30 October, approving the revised text of the Law on the Basic Statute of the Public Employee). In the case of subsidies, they are regulated as provisions (article 17 of Law 38/2003, of 17 November, General Law on Subsidies).

adherence, acceptance and voluntary subscription by the participant to unilateral administrative acts of the Administration in which terms, conditions and regime of participation in the sandbox are established. This is the case with the "Protocols" in the financial sandbox.[66] The AI Act requires the existence of a "specific plan agreed between the providers or potential providers and the competent authority" (Art. 57.5). These protocols or specific plans are linked to the call's terms and detail the specific regime of obligations.

VI. Exceptionality of the legal regime and liability of participants. Theory, reality and Regulation

As discussed below, there is a complex interplay between theory and reality regarding the exceptions or singularities involved in a sandbox, particularly an AI sandbox. On this basis, regulation in AI Act is analysed. In theory, *a regulatory sandbox implies various formulas of exceptionality in the validity, application or enforceability of existing law*.[67] This issue has been the subject of constitutional regulation in France (Art. 72), as well as case law of the German TC[68] and the CJEU for the EU in relation to sandboxes.[69] It must therefore be regulated which legislative provisions can be waived, exempted, not pursued or repealed, and, if necessary, replaced by decisions of the sandbox authorities. The nature of the liabilities involved should be specified (civil, administrative, etc.) and whether they are punitive or criminal.

From the principle of competence, it should be clarified whether the sandbox implies an alteration of competences or the specific empowerment,

[66] Article 3 of Law 7/2020 of 13 November on the digital transformation of the financial system regulates the "Protocol", which is the document containing the terms under which the tests will be carried out. It will be signed by the promoter and the supervisory authority or authorities that are competent for the subject matter of the project.

[67] As recalled by BMWi, *Making Space for Innovation… cit.* pp. 82 and following. The legal-normative typology can also be found in Conseil d'État, *Les expérimentations…* cit.

[68] Particularly, the French Constitutional Council's regulations and decisions, as well as the significant actions of the Council of State, play a significant role.

In Germany, in addition to BMWi, *Making Space for Innovation… cit.,* the rulings of the Federal Constitutional Court should be taken into account. First of all, it has given them constitutional status in its decision of 16 June 1981 (the so-called third broadcasting decision), also, the decision of the Federal Constitutional Court of 24 March 1987, the so-called fifth broadcasting decision, "fünfte Rundfunkentscheidung").

[69] In particular, CJEU, Opinion of the Advocate General in case C-127/07. As long as the experimental laws are transitional in nature and the judgment is based on objective criteria, the inherent need for differentiation of a sandbox is compatible with the principle of equal treatment.

attribution or delegation of a regulatory authority or body, which should be established by a rule of the same rank that regulates the altered competence.[70] It should also be specified which authorities would be affected by the exceptions or particularities (data protection, Artificial Intelligence, sectoral, etc.). Ranchordas[71] cautioned that in order to prevent fragmentation, the scope of action of national law should be regulated by EU law itself. In principle, the exceptionality of a rule must be established in a rule of higher or equal rank.[72]

While in theory the exceptions or singularities involved in a sandbox should be clearly regulated, *in practice existing sandboxes do not clearly address sanctioning exemptions*. In the known experiences, this issue is not clearly addressed. For example, the 2015 UK financial sandbox was aware of the lack of regulatory coverage in EU law for exemptions, and it was stated that "The government could consider changing the exemption conditions in FSMA to make it easier for the FCA to waive the rules for a firm inside the sandbox".[73]

In Spain, it is difficult to find legal coverage for a regulatory exemption or non-application of regulations. Article 15.4 of Law 7/2020 of 13 November, which regulates the financial sandbox, establishes that, if the participant follows the sandbox law and protocols, there is an exemption or exoneration, but only regarding their participation in the sandbox. However, the authorities retain all their powers and liability rules for damages.[74]

[70] See Guño, A., *Sandbox Regulatorio… cit.* p. 20. There are technical and procedural provisions in which the law provides for the authorities themselves to define certain rules on the matter. It is not uncommon to see in different regulations that the legislative body gives agencies or regulatory bodies the power to generate some specific rules, considering the knowledge and experience they have in the matter. In this way, the authority may experiment by generating a new circular or other documentation to define new provisions, as long as it has the legal power to do so. In this case, it is important to consider the phenomenon of discretionality within regulatory agencies that has been extensively analysed in US academia. Likewise, given the impossibility for Congress to provide all the specific elements required to apply a rule, there is room for interpretation and even for deciding when to apply or not to apply a rule.

[71] Ranchordás, S., "Experimental Regulations for AI…" *cit.*

[72] Guño, A., *Regulatory Sandbox… cit.* "specific rules can be generated that allow experimentation with other rules that respect the corresponding regulatory hierarchies. In this way, if what is to be experimented with is indicated in one law, it is because another law allows it. In general terms, it would not be admissible for a lower-ranking law to allow experimentation with the content of a higher-ranking law".

[73] Financial Conduct Authority, *Regulatory sandbox*, 2015, pp. 14-15, https://www.fca.org.uk/publication/research/regulatory-sandbox.pdf

[74] In the financial sandbox, the main exemption during testing is that the promoters are not subject to the financial regulations that impose an administrative authorisation (art. 4. 2, Law 7/2020). As regards compliance with the legal regime, there is an 'exoneration' and 'ex-

In the field of AI, no rule providing for a general exemption has been found. Despite the fact that the essence of a sandbox is to involve special regulation and exemption, institutions are wary of seeing their control and sanctioning powers limited, especially if they do not participate in the sandbox. In the Spanish case of the AI sandbox, data protection seems obvious, and the GDPR does not contemplate the possibility of exemptions. Royal Decree 817/2023 of 8 November does not provide for any exceptionality (Art. 4 on "Legal regime").

In the cases of the UK, French or Norwegian AI sandboxes, it is stated that it is not possible to exempt compliance with data protection regulations. Informally, in their guides, websites and publications, eclectic formulas are used to convey that irregularities will not be prosecuted, but rather the opposite. Of particular interest is the ICO's informal solution,[75], which points out

emption' in respect of specific activities for participation in the sandbox, and 'within the limits of the pilot project' (art. 4.3, Law 7/2020).

In other words, it is assumed that if the sandbox law and protocols are followed, there is exemption or exoneration. If they do not follow protocols, the regulations do apply, and in these cases, liability is specifically underlined for those who "also infringe management or disciplinary rules" (art. 15).

However, among those who do follow the law and the sandbox protocols, their exemption or exemption is limited to their activities related to their specific participation in the sandbox. But "In no case shall this exemption extend to ordinary activities outside the controlled test space" (Art. 4.3). However, in these cases, although there is no exemption, a "weighing of the principle of proportionality" is foreseen (Art. 4.3 and 19).

However, the obligation and the responsibility to impose penalties on those who "also infringe planning or disciplinary rules" (Art. 15) is maintained.

It should also be recalled that the authorities retain their competences (Art. 2). Rules on liability for damages are generally maintained (Art. 12).

Furthermore, the "monitors" assume no liability for non-compliance by the participants (art. 3).

[75] https://ico.org.uk/media/for-organisations/documents/2618111/sandbox-terms-and-conditions.pdf The UK ICO in its "Terms and Conditions" states that "You agree that you remain responsible for your compliance, and the compliance of your proposed Innovation, with all legal and regulatory obligations, whether in respect of data protection law or otherwise". (1.6). And that "Acceptance into the safe harbour does not preclude regulatory action by us or any other competent data protection authority or any other regulatory body or authority. Comments do not affect the rights conferred on third parties (such as your customers), nor do they bind any court, and may not reflect the views of any other data protection authority." (1.10). Now, in the same document, in response to "What happens if we encounter a personal data breach while our product is in the sandbox?" as well as stating that "we expect you to report it to the ICO within 72 hours, in accordance with the UK GDPR requirement" it is expressly stated that "Although the ICO will consider the breach in accordance with our standard procedures, we are very unlikely to take enforcement action if you are complying with

the obligation of communication of possible irregularities by the participating entity, while indicating the limited possibility of sanctioning action. As regards the possibility of non-compliance with sectoral legislation, it states that it will not take proactive action to report possible non-compliance. It is important to remember that the sandbox information and participation guides, despite not being regulatory in character, can still have a legal significance. Consequently, they can be used to assess the specific culpability of the entity participating that is involved in a sandbox. Whoever follows the indications expressly formulated by the various channels and can prove it, can hardly be considered to have engaged in culpable conduct. Euphemisms are also found in Norway, where the *Datatilsynet* in 2021 stated that "The sandbox cannot grant exemptions from the regulations. *The Data Protection Authority does not intend to initiate corrective measures during an organisation's participation in the sandbox. The focus will be on helping participants to comply with existing regulations.*"[76] More restrained is France's CNIL. On its website it states that "This sandbox cannot lead to the lifting of regulatory restrictions, even temporarily, because the European texts on data protection (GDPR) do not provide for an exemption on this ground. *However, it does have an experimental, testing vocation*, to resolve a difficulty or uncertainty, identified in collaboration with the project leader".[77]

Analysing the issue in the light of the AI Act, it can be seen that it has varied considerably in its handling. The first version and the last version adopted are contrary regarding the possible exoneration or exemption from punitive liability. The initial version (Art. 53.3) expressly denied exoneration or failure to act on the part of the competent authorities.[78] However, in the final version, the starting point is that sandboxes "shall not affect the supervisory or corrective powers of the competent authorities supervising the sandboxes" (Art. 57.11). However, an exemption from sanctioning responsibilities is allowed as long as providers "observe the specific plan and the terms and

the terms of your letter of entry to the sandbox". It is also stated that "the test environment team will not proactively assess the compliance of your organisation or your processes in general. If we identify a reportable breach during the course of the Sandbox […] we will advise you to report it to the ICO".

[76] https://www.datatilsynet.no/en/regulations-and-tools/sandbox-for-artificial-intelligence/framework-for-the-regulatory-sandbox/what-are-the-relevant-regulations/

[77] https://www.cnil.fr/fr/un-bac-sable-rgpd-pour-accompagner-des-projets-innovants-dans-le-domaine-de-la-sante-numerique

[78] "3. Controlled AI test sites shall not affect the supervisory and corrective powers of the competent authorities. Any significant risk to health, safety, security and fundamental rights identified during the development and testing process of these systems shall entail immediate mitigation and, failing that, suspension of the development and testing process until such mitigation takes place.

conditions for their participation and follow in good faith the guidance given by the national competent authority, no administrative fines shall be imposed by the authorities for infringements of this Regulation." (Art. 57. 12). This exemption does not extend to the application of other applicable sectoral law, such as data protection. Nevertheless, data protection compliance may be exempted if "other competent authorities responsible for other legislation have been actively involved in the supervision of the AI system in the sandbox and have provided guidance for compliance" (Art. 57. 12º).

In addition to the exceptional regime that the sandbox may entail, it is necessary to focus on *the particular regime of liability for damages*. During the testing phase of a pilot project, damage to third parties may occur. It is therefore necessary to define how liability will be allocated between the various participants in the process. Currently, in the European Union, there is no specific legislation regulating liability for damage caused by Artificial Intelligence systems.[79] The AI Act affirms the liability of providers (Art. 57. 12) "under Union and national law and national liability law for any damage inflicted on third parties as a result of the experimentation taking place in the sandbox ". This is without prejudice to the liability for penalties already mentioned. In Spain, liability for damages suffered by participants is regulated, for example, in Article 12.1 of the Financial Sandbox Law 7/2020 of 13 November,[80] or in the not yet approved Sustainable Mobility Law.[81] Moreover, the majority of the sandboxes in Spain that are either planned or regulated allow for the exclusion of liability for the public authorities that are involved in the development of the test pilot,[82] also in the ill-fated AI sandbox (Art. 17 of Royal Decree

[79] In this regard, Expert Group on Liability and New Technologies of the European Commission, *Liability for Artificial Intelligence and other emerging digital technologies*, 2019. https://op.europa.eu/en/publication-detail/-/publication/1c5e30be-1197-11ea-8c1f-01aa75ed71a1/language-en Similarly, since 2022 there is a Directive on liability in the field of AI, COM/2022/496 final in the pipeline on the subject, on which the literature is very abundant, Atienza Navarro, M. L., *Daños causados por inteligencia artificial y responsabilidad civil,* Atelier, Madrid, 2022.

[80] "The liability for damage suffered by participants as a result of their participation in the events shall be borne exclusively by the promoter when it is caused by his failure to comply with the protocol, when it arises from risks of which he was not informed or when there is fault or negligence on his part. In the event of damage resulting from technical or human error during the course of the events, the liability shall also be borne by the promoter".

[81] It is based on the presumption of liability of the promoter, although the concurrence of faults is expressly foreseen and the existence of a system of guarantees to be established in the protocol is also affirmed.

[82] Thus, in article 12.1, Law 7/2020 financial sandbox: "The authorities that intervene during the development of the tests shall not be liable for any possible damages that may

817/2023 of 8 November).[83] Despite the regulated exemptions, it would be necessary to follow case law.[84] Another element to be regulated is the system of guarantees for participants, which must be presented before the test pilot or sandbox begins. However, the AI Act does not regulate these issues.

VII. Sandbox authorities, selection of participants, duration, development, and obtaining conformity assessment

1. The competent sandbox authority and its cooperation with other national and European authorities

Under the AI Act, it is the competent authorities that "establish" a sandbox for AI (Art. 57. 1). These authorities may be at national level, although the possibility of "additional sandboxes at regional or local level or jointly with the competent authorities of other Member States" is also envisaged (Art. 57.2). The AI Act does not specify who should be the authority for a sandbox and it does not have to be the market surveillance authority, unlike what happens with tests under normal conditions. In the EU context, such sandboxes would be established by the European Data Protection Supervisor (Art. 57.3). The authorities establishing them are responsible for providing "sufficient resources" to comply with the AI Act.

The establishment of an AI sandbox would be carried out under the framework of the AI Act and national legislation, in Spain through a regulation and concretised through specific bases. In addition, there would necessarily be a "Specific Plan" agreed by the parties. Adequate resourcing must be ordered normatively and institutionally, as well as provided for in the budget.

Authorities establishing the sandbox should provide "supervision and support within the AI regulatory sandbox" with a view to identify risks, support "testing and mitigation measures and their effectiveness" (Art. 57.6) and provide "guidance on regulatory expectations and how to fulfil the require-

arise". Or in Additional Provision 23, Law 24/2013, of 26 December, on the Electricity Sector or in the Draft Bill on Sustainable Mobility, article 71.5.

[83] "Both the participating AI provider and, where applicable, the participating user shall be liable for damages suffered by any person as a result of the application of the Artificial Intelligence system in the context of the controlled test environment, provided that such damages result from a breach or where there is fault, negligence or wilful misconduct on their part".

[84] Thus, several judgments consider the financial liability of the administration by applying *culpa in vigilando*, related to cases of defective "inspection or supervision"*:* STSJ Aragón 15-2-1999; STS 25-1-1992; STC 112/2018; SAN 24-6-2019.

ments and obligations" of the AI Act (Art. 57.7). The supervision of AI systems in controlled space should cover their development, training, testing and validation prior to their introduction on the market or putting into service, as well as the concept of "substantial modification" and its materialisation (Consid. 139).

In the organisation of a sandbox, governance must be established to include data protection or other authorities, which are "linked to the operation" of the sandbox (Art. 57. 10º). The competent sandbox authorities "shall have the power to temporarily or permanently suspend the testing process, or the participation in the sandbox if no effective mitigation is possible, and shall inform the AI Office of such decision" (Art. 57. 11).

The AI Act provides for cooperation between national AI sandbox authorities and the European AI governance framework. AI sandboxes "shall be designed and implemented in such a way as to facilitate, where appropriate, cross-border cooperation between competent national authorities" (Art. 57. 13). In addition, "the competent national authorities shall coordinate their activities and cooperate within the framework of the Committee" (Art. 57. 14). They "shall inform the AI Office and the Committee of the establishment of a controlled evidence area and may request support and guidance from them" (Art. 57. 15).

A sandbox register by the AI Office is also foreseen to facilitate interaction and cooperation (Art. 57. 15º) and a system of annual reports by national authorities to the AI Office and the Committee after their completion, which will be made public (Art. 57.16º). All information will be managed through a "single, dedicated interface" coordinated by the Commission (Art. 57.17º).

2. Selection and admission of participants and duration of the sandbox

A relevant aspect is the definition of eligible participants for the sandbox, the admission requirements, and the fundamental procedure. The regulation needs to determine precisely the sector involved, objectives and conditions of application.[85] The OECD recalls that sandboxes are selective due to resource constraints, and participants are selected on the basis of eligibility criteria. Of the 63 applicants to the 2019 ICO UK sandbox, only ten were selected based on clearly determined criteria.[86]

[85] Conseil d'État, *Les expérimentations…* cit. However, the law or decree must define with sufficient precision the objective of the experimentation and the conditions for its application (CC, no. 2004-503 DC of 12 August 2004 or CC, no. 2019-778 DC of 21 March 2019).

[86] OECD, *Regulatory sandboxes… cit.* p. 25. Thus in UK ICO (2019), *Information Commissioner's Office Regulatory Sandbox*, https://ico.org.uk/for-organisations/regulatory-sandbox/.

The selection elements of a sandbox may be similar to public procurement procedures[87] or other processes for the attribution of advantages or subsidies. The AI Act takes this issue into account and implementing acts adopted to avoid fragmentation in the EU must ensure that any provider has access with transparent and fair eligibility and selection criteria (Art. 58.2a). Broad and equal access" will be allowed, with the possibility to bid "in partnership with deployers and other relevant third parties" (Art. 58.2a). Equal advantages are also foreseen for SMEs and startups (Art. 58.2 d) and access by "other relevant actors in the AI ecosystem […] to allow and facilitate co-operation with the public and private sectors". Examples include "notified bodies and standardisation organisations, SMEs, including start-ups, enterprises, innovators, testing and experimentation facilities, research and experimentation labs and European Digital Innovation Hubs, centres of excellence and individual researchers" (Art. 58.2 f).[88]

As regards the *duration of the sandbox*, this is an essential element and must be fixed by the act establishing the sandbox.[89] Possible derogations, exemptions or non-applications must be transitional[90] and their duration cannot be left to the free decision of the administration. The question must be asked: "How long will it take to achieve the objectives of the regulatory sandbox?[91] The duration should be appropriate to the nature of the test on objective criteria, with sufficient time for representative and valid tests, but not excessively long or permanent. The timeframe may be determined by dates, months from initiation, or determinable circumstances. It may also be determinable by the sandbox authority on the basis of certain circumstances. The possibility of regulating possible extensions of the time limit based on decisions of authorities or certain circumstances should not be excluded, given the element of innovation and lack of knowledge.[92]

[87] BMWi, *Making Space for Innovation… cit.*

[88] Examples include "notified bodies and standardisation bodies, SMEs, including start-ups, enterprises, innovators, testing and experimentation facilities, research and experimentation laboratories and European Digital Innovation Centres, centres of excellence and researchers".

[89] The French Constitutional Council has stated that the limitation of its duration is inherent to experimentation: it must be fixed by the act instituting it. When the legislator decides on an experiment, it cannot leave it to the regulatory authority to set the time limit (CC, no. 2009-584 DC of 16 July 2009) See BMWi, *Making Space for Innovation… cit.* para. 3.

[90] In particular, CJEU, Opinion of the Advocate General in case C-127/07, EU:C:2008:728). It is stated that the inherent need for differentiation of a sandbox is compatible with the principle of equal treatment as long as the experimental laws are transitional in nature and the judgment is made according to objective criteria.

[91] BMWi, *Making Space for Innovation… cit…* Section 3 Design, pp. 80 ff.

[92] Some experimentation clauses also allow for a later extension of the term. The option

The AI Act states that Commission implementing acts specifying EU-wide sandbox rules and criteria shall ensure "that participation in the controlled AI test area is limited to a period that is appropriate to the complexity and scale of the project, and that may be extended by the national competent authority" (Art. 58.2 h).

3. Developing, completing and achieving a sandbox compliance assessment

As noted above, the development of the sandbox implies that the "competent authorities shall provide, as appropriate, guidance, supervision and support" (Art. 57.6), and offer "guidance on regulatory expectations and how to meet the requirements" (Art. 57.7). The completion of the sandbox should involve actions for general and specific evaluation and feedback for each participant. This can be articulated through final reports, memories or conclusions. It is possible to regulate some minimum elements of these documents and, in particular, their publicity regime. In this respect, the AI Act provides that, if requested by the sandbox participant, "the competent authority shall provide written evidence of successfully completed activities" and "provide an exit report" and "corresponding learning outcomes". This may be relevant for "demonstrating its compliance with this Regulation through the conformity assessment process or relevant market surveillance activities" (Art. 57.7). This must also be ensured in the Commission's implementing acts (Art. 58. 2 e).

In particular, it states that "the Commission and the Committee shall be authorised to access the exit reports and shall take them into account, as appropriate, when exercising their tasks under this Regulation ". For sandbox exit reports to be made public, the participants must give their consent (Art. 57.8).

VIII. Data protection in the context of an Artificial Intelligence sandbox

The AI Act has devoted special attention to data protection in sandboxes, in particular to the "further processing of personal data for the development of certain AI systems in the public interest in the controlled AI test space" (Art. 59).

to extend the project can be useful, especially in the case of experimentation clauses with short deadlines, to increase the degree of flexibility in the initial testing phase.

Any use of Artificial Intelligence involving data processing must comply with data protection regulations and, among other aspects, have a legitimate basis for processing personal data (Art. 6 GDPR). AI systems seeking to engage in the sandbox possess personal data for which, they have, in principle, legitimate authorization based on permissions, contract executions, adequate legal regulation, etc. However, processing data for the purposes of the sandbox may be deemed incompatible, so they would not be able to process data in this context. This is the area in which the AI Act plans to operate.

The AI Act allows in a sandbox and "solely for the purpose of developing, training and testing certain AI systems" the processing of data collected lawfully for other purposes. The AI Act becomes the legal basis for doing so, provided that "all of the following conditions are met". These are ten paragraphs of requirements that must be fully met in order to legitimise the processing of data in the sandbox. Among such conditions, the purposes of processing are defined (essential public interest purposes or processing data for the fulfilment of AI Act obligations), that there are effective supervision mechanisms, that the data remain under the control of providers, that there are adequate technical and organisational measures and that the data are then deleted, that there is functionally separate, isolated and protected data processing, that data cannot be transferred and cannot leave the sandbox, that no decisions affecting data subjects are generated, that there is a record of logs, as well as a full and detailed description of the process, or, finally, that a short summary of the AI project is published. It is foreseen that there may be some specific regulation "to develop, test or train innovative AI systems or any other legal basis" (art. 59. 3º).

Well, this is a very detailed and demanding regulation, but it is only limited to legitimising data processing of legal origin in order to facilitate access to the sandbox. It also suggests specific guidelines for data processing within this context. Despite the detailed rules, Article 59 does not facilitate participation in a sandbox from a data protection perspective. This is due to the potential fear of participants being subject to a thorough data protection compliance check.

The difficulties of ensuring comprehensive regulatory compliance in such an innovative context should not be overlooked. Many of the various high-risk AI systems which are data processing systems, currently lack a clear basis for legitimisation or regulation. It is not simple for those who want to participate to ensure full compliance with these complex regulations. As a result, participating in a sandbox can mean *exposing oneself*, even *putting oneself in a mousetrap* in the eyes of data protection authorities. Also, the AI Act does not address some relevant data protection issues.

The participating provider's required data protection compliance information could potentially hinder access to the sandbox. In some cases, the sandbox authority may not be the data protection authority, which adds complexity. An intense requirement to demonstrate data protection compliance for access and participation can be a clear inhibitor. And the focus of the sandbox need not be data protection compliance.

In the case of the AI sandbox in Spain, Article 16 of Royal Decree 817/2023 of 8 November states that "AI providers and users participating in the controlled test environment shall comply with the provisions of" data protection regulations. The annexes include a "Responsible declaration of compliance with the principle of proactive responsibility for data protection" (Annex IV). It states that they have adopted proactive accountability measures and that they may be required to provide supporting documentation. It is also stated that "non-compliance with these regulations will result in the definitive termination of the tests". Annex V specifies the documentation "which may be required" in ten points. To a large extent, it is clear that the intention is not to set up a system to control compliance with data protection regulations in the sandbox, but rather to protect itself from problems that may arise in this regard and from a minimum of responsibility in this area.

It is important to note that the sandbox can be a place to identify data protection breaches. As a matter of principle, the competence of the sectoral authorities is not altered, so the data protection authority will maintain its full supervisory powers over sandbox participants (Art. 57. 11º). There would be no exemption from sanctions if the data protection authority does not actively participate in the sandbox (Art. 57. 12).

Finally, another element that could be taken into account is that if a relevant data protection incident occurs in the sandbox, it may be mandatory to communicate it to the data protection authority, which may be another inhibiting element for participation. In the -maligned- AI sandbox in Spain, Article 15 of Royal Decree 817/2023 of 8 November imposes the obligation to communicate to the sandbox authority "any serious incident in the systems that could constitute a breach of the legislation in force". In addition, for AI systems that "are subject to other specific legislation, the competent body shall transfer the communication to the competent sectoral authorities, and it shall be up to the sectoral authorities to take such measures as they deem appropriate".

These problems are highly pertinent in practice and remain unregulated in the AI Act. National legislation instituting the sandbox could alleviate several of the aforementioned issues.

IX. Testing under high-risk Artificial Intelligence system conditions

In the initial proposal of the AI Act by the Commission there was no reference to "Testing of high-risk AI systems under real conditions outside controlled AI test areas". These were introduced in the internal Council versions during the French Presidency in the first half of 2022 and have finally been regulated in Articles 61 and 62.

Recital 141 states that "it is important that providers or prospective providers of such systems may also benefit from a specific regime for testing those systems in real world conditions, without participating in an AI regulatory sandbox ". This should be done with "appropriate and sufficient safeguards and conditions", such as "informed consent of natural persons", which is distinct from data protection consent. It is also intended to "minimise risks and allow oversight by competent authorities". This requires submitting to the authority "a plan of the test under real conditions" and that providers register the test in specific sections of the EU database. It also requires a "written agreement defining the roles and responsibilities of potential providers and those responsible for deployment, and effective supervision by competent personnel involved in the real-world test". A number of "additional safeguards" are foreseen[93] and some relating to data transfer.

It should be recalled that testing under real conditions can also take place within a controlled test environment (Art. 57. 7º).[94] The legal regime that would be applicable in these instances would be that of the controlled environment., but the authorities "shall specifically agree on the conditions", including "appropriate safeguards with the participants, with a view to protecting fundamental rights, health and safety" (Art. 58.4).

Regarding limitations and conditions for real-world testing. Real-world testing is limited to "y providers or prospective providers of high-risk AI systems listed in Annex III" (Art. 60. 1º). In some cases, such testing will be close to research. This may come as a surprise, as the AI Act does not apply to "AI systems or models, including their output results, developed and put into service specifically for the sole purpose of scientific research and development". However, research with a view to developing AI systems for commercialisation may be considered applicable to the "potential" providers referred to in Article 60. Hence the references to the consent of those affected and the

[93] "to ensure that it is possible to effectively reverse and discard the predictions, recommendations or decisions of the AI system and that personal data are protected and deleted when subjects withdraw their consent to participate in the test".

[94] Thus, Article 57.7 states that "Such controlled test sites may include tests under actual

overlap with “any ethical review” required for such testing, as in the field of research (Art. 60(3)).

In terms of requirements and responsibilities of providers. Possible specific legislation is foreseen for the testing of high-risk AI systems in Annex I, that is, products of a certain danger levels subject to third-party conformity assessment that incorporate AI systems. The provider must be established in the Union or have appointed a legal representative (Art. 60(4)(d)). It is possible to organise the tests “in cooperation with one or more deployers”. In such cases, they must be well informed and there must be an agreement between the processor and the data protection officers (Art. 60. 4 h). The provider and the deployers must effectively supervise the tests with qualified, trained and authoritative personnel (Art. 60.4j). They must also report any serious incidents, take action or suspend testing, and have a procedure for the rapid recovery of the AI system (Art. 60. 7). If testing is suspended or terminated, they must notify the authority (art. 60. 8º). “The provider or potential provider shall be liable […] for any damage” (art. 60. 8º).

Regarding the procedure and duration of the tests. The tests may be carried out “at any time before the placing on the market or the putting into service of the AI system on their own or in partnership with one or more deployers or prospective deployers.”. (Art. 60. 2º). The duration shall be as long as necessary and not longer than six months, extendable for a further six months with notification and explanation to the authority (Art. 60. 4º f).

The Test Plan is essential for establishing their legal regime, and the Commission details these plans in an implementing act (art. 60. 1º). Whoever intends to carry out the test submits the plan to the State supervisory authority, which must approve it. In principle, it is considered to be approved if there is no response within thirty days (art. 60. 4º b). The market surveillance authorities shall have effective powers to request information and may carry out unannounced inspections at a distance or *on site* and control the performance of the tests (art. 60. 6º).

Regarding the protection of the individuals involved in the tests, the applicable regime may align with the cases under investigation, thereby ensuring a certain uniformity in their treatment. Adequate protection” of the persons concerned is foreseen if they are “vulnerable groups due to their age or disability” (Art. 60. 4 g). It must be ensured that “predictions, recommendations or decisions of the AI system can be effectively reversed and discarded” (art. 60. 4º k). Furthermore, data subjects may “withdraw from the tests at any time by withdrawing their informed consent and request the immediate and

supervised conditions within them”.

permanent deletion of their personal data", without prejudice (art. 60. 5º). The subjects of these tests must give an informed consent regulated in Article 61, different from the data protection consent (Consid. 141). This precept details the information to be provided on the nature and objectives of the tests, the conditions, duration, rights and guarantees, including the right to abandon the tests, the possibility of requesting the reversal or discarding of the predictions, recommendations or decisions of the system, and information on the unique identification number of the tests (Art. 61. 1º). With regard to personal data, personal data will only be transferred to third countries if the requirements of the GDPR are met.

X. SMEs, start-ups and micro-enterprises in the Regulation

Finally, it remains to refer to Articles 62 and 63, which deal with measures targeted at SMEs and start-ups, as well as exemptions for micro-enterprises. The AI Act aims to give special consideration to the interests of SMEs, including start-ups, which are providers or responsible for the deployment of AI systems (Consid. 143). It should be recalled that the Commission is to adopt "Guidelines" on the implementation of the AI Act and "shall pay particular attention to the needs of SMEs, including start-ups, local public authorities and the sectors most affected by this Regulation". The speciality of these companies will also be considered in the application of the sanctioning regime (art. 99. 1º and 6º).

As mentioned above, priority or easy access to sandboxes is foreseen. Also Article 62 states that Member States shall "organise specific awareness-raising and training activities on the application of this Regulation tailored to [their] needs", "utilise existing dedicated channels and where appropriate, establish new ones for communication […] to provide advice and respond to queries about the implementation of this Regulation" and "facilitate the participation […] in the standardisation development process. ". In setting fees for conformity assessment, "specific interests and needs shall be taken into account". It highlights the "advisory forum" regulated in Article 67 to integrate the interests of SMEs.

It should be recalled that SMEs and startups can comply in a simplified manner with the technical documentation (art. 11. 1º). Despite the absence of explicit reference to SMEs and startups, this precept stipulates that the AI Office will have standardised templates, a single information platform, appropriate communication campaigns and will assess and promote the convergence of best practices in public procurement procedures in relation to

AI systems. It should be recalled that under Article 95.4 “the AI Office and the Member States shall take into account the specific interests and needs of SMEs, including start-ups, when encouraging and facilitating the drawing up of codes of conduct “.

For the case of “micro-enterprises”[95] Recital 146 and Article 63 seek to make compliance with the regulation, altough in a very limited manner. Basically, the favourable treatment is limited to the fact that they may comply “in a simplified manner” with some elements of the quality management system required by Article 17. It is foreseen that “the Commission should develop guidelines to specify the elements of the quality management system to be fulfilled in this simplified manner by microenterprises” (Recital 146). However, the exceptionality for micro-enterprises is very limited; for the avoidance of doubt, it is provided that such operators are not “exempted from fulfilling any other requirements or obligations laid down in this Regulation, including those established in Articles 9, 10, 11, 12, 13, 14, 15, 72 and 73.” (Art. 63. 2).

XI. Conclusions and summing up

This study has focused on sandboxes or controlled spaces, as well as on real-world testing of Artificial Intelligence systems. Measures for SMEs, start-ups and micro-enterprises have also been addressed, as they are included in this Chapter V of the AI Act.

The AI Act aims to facilitate innovation through these relatively new tools, which still raise a number of misgivings, especially among jurists In fact, the AI Act imposes an obligation on each Member State to establish at least one sandbox within two years of the entry into force of the AI Act. In contrast to the existing regulatory dispersion surrounding sandboxes, the AI Act establishes a homogeneous regulatory framework for the entire EU, while allowing for flexibility and controlled experimentation, which are crucial for the development of emerging technologies.

It has been observed how sandboxes, initially linked to sectors such as Fintech, have proven to be effective tools for testing and validating technological innovations in a safe and controlled environment. There is a certain

[95] They are defined in Article 1 of the Commission Recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises: “Any entity, regardless of its legal form, engaged in an economic activity shall be considered to be an enterprise. In particular, entities engaged in artisanal activities or other activity on an individual or family basis, partnerships and associations that regularly engage in economic activities will be considered as enterprises”.

terminological variety that, in a way, the AI Act resolves with its sandbox concept for its own normative purposes. The international experiences of AI sandboxes have been analysed. Those of the UK, Norway and France show that their focus has been on data protection compliance, although the AI Act will promote other perspectives possibly of greater interest, perspectives, in any case, compatible with data protection. International collaboration and the creation of common databases of relevant use cases, as contemplated by the AI Act, are essential for sharing knowledge and improving regulatory practices across the EU. Likewise, the numerous advantages of sandboxes have been highlighted from different perspectives: fostering innovation and competitiveness, improving legal certainty and facilitating market access for SMEs and startups. These environments allow for more adaptive and dynamic regulation, attracting investment and improving the global competitiveness of the EU in the field of AI.

Based on the above, the focus has been on the regulatory framework of a sandbox. In order to prevent regulatory fragmentation among the Member States, the AI Act regulates this regulatory framework. Indeed, the European Commission will adopt implementing acts including common elements detailed in Article 58 for the implementation of sandboxes. This EU framework does not exclude the possibility that national regulations may establish sandboxes outside AI. Furthermore, as long as they do not violate the AI Act, States mantain the ability to regulate sandboxes. It has been explained how there is a basic legal framework in Spain and that each specific sandbox is established through regulations (such as the ill-fated AI sandbox planned for 2022 in Spain). This regulation encompasses the bases of the call, protocols or agreements, and the "Plan" of the sandbox, which establishes the rules for the AI Act between the authority and the participants.

An element that is theoretically essential in a sandbox has also been discussed: the exceptional nature of the legal regime and the liability of participants. Theoretically, it is possible in sandboxes to temporarily exempt participants from certain regulatory obligations and sanctioning responsibilities. However, the AI sandbox experiences that have occuredhave resorted to various subterfuges to address the issue. The AI Act stipulates that providers who adhere to the guidance provided by the competent authority in good faith will not be subject to administrative fines for violations of the AI Act during their participation in the sandbox. On the other hand, providers remain liable for any damage caused to third parties during sandbox experimentation. In any case, data authorities, for example, will still be able to apply their regulatory regime and sanctions, unless they participate as a sandbox authority. In this respect, the regulation of AI Act has been analysed with regard to competent authorities at national, but also regional or local level, which should provide continuous oversight and support to participants. Important precautions and requirements regarding the admission and selection of participants have also been highlighted, which must be based on transparent and equitable eligibility criteria, with priority access for SMEs and startups. The duration of the sandboxes is initially set for six months, which can be extended. There should be monitoring and collaborative learning during the sandbox and at the end, the competent authority should provide an exit report documenting the activities carried out and the learning gained, which, as a novelty, may be relevant to demonstrate compliance with the AI Act and facilitate the conformity assessment.

The AI Act also pays attention to the processing of data in AI sandboxes. The regulation focuses, perhaps too much, on the legitimisation of the processing of personal data by providers and deployers. Very precise, perhaps excessive, requirements are imposed to consider that participation in the sandbox is not an incompatible treatment. Nevertheless, various aspects of data protection that I believe create an inhibiting effect on participation in sandboxes are not addressed, due to the fear of exhaustive control by data protection regulations. National legislation should mitigate these negative effects to avoid discouraging participation in sandboxes.

Chapter V also regulates innovative real-world testing of high-risk AI systems outside sandboxes. This is a reality that can be very close to AI research, hence can be performed by "potential providers". Such testing under real-world conditions is essential to assess the feasibility and safety of AI systems, but must be conducted under strict conditions to protect the fundamental rights of the individuals concerned. The AI Act regulation sets out a clear, albeit complex, framework for these testing, highlighting the need for a detailed test plan and effective oversight by competent authorities. In addition, and due to the aforementioned proximity to the field of research, informed consent of those affected is regulated with a series of guarantees.

Finally, the focus on SMEs, start-ups and micro-enterprises in the AI Act is addressed. Support measures are foreseen, such as priority access to sandboxes and the possibility to comply with regulatory requirements in a simplified manner. However, the flexibility of AI Act compliance for these companies is limited and they must comply with most AI Act obligations. The European Commission's guidelines will be indispensable in determining how they can simplify specific aspects of the quality management system.

Sandboxes, or controlled spaces, and real-world testing of Artificial Intelligence systems are crucial tools for technological innovation and for learning how to implement the AI Act itself.

AI Act provides a homogenous regulatory framework across the EU, allowing the necessary flexibility to experiment and develop emerging technologies. AI Act ensures a safe and controlled environment for experimentation. Dozens of sandbox experiments will take place in the coming years, and a common learning and experience of the application of AI Act itself should be generated. It will also be seen whether the regulation of testing under real conditions is adequate and effectively enables research and innovation. Similarly, time will also tell whether there is a need for more specific provisions for small businesses in the EU. The success of these initiatives will depend on continued collaboration between national and European authorities, and the ability to adapt regulation to the changing needs of the sector.

GOVERNANCE AND OVERSIGHT OF THE ARTIFICIAL INTELLIGENCE ACT: MARKET SURVEILLANCE AUTHORITIES, THE COMMISSION AND THE VARIOUS ENTITIES

Juan Carlos Hernández Peña
Senior Lecturer in Administrative Law
University of Navarra

I. Introduction

In order to establish a reliable AI ecosystem that combines safety, health, protection of fundamental rights and responsible innovation, it is necessary to establish a comprehensive and effective system of governance, as well as to define the procedures to be followed in cases of non-compliance.

This chapter explores these issues in greater depth, starting with an analysis of the multi-level governance model delimited by Chapter VII of the Act. To this end, it develops the competences attributed to the European Commission and the Member States, as well as to the main administrative bodies of the new institution: the AI Office; the European Committee on Artificial Intelligence; the Advisory Forum; and the Scientific Panel of Independent Experts.

Subsequently, we will refer to regulated entities which, although common to other sectoral areas, the Regulation mandates their establishment or gives them competences in the field of Artificial Intelligence. This is the case of market surveillance authorities or notified bodies, already provided for by the general product safety regulation, as well as the European Centre for Algorithmic Transparency, created under the Digital Services Regulation.

The final sections of this chapter will be devoted to an analysis of the market surveillance measures to ensure the protection of the public interest in case of doubts about the conformity of an AI system. These are the procedures for market surveillance and control of AI systems in the EU, which are largely based on and adapt those established by Chapter III of Regulation 765/2008[1], as well as Article 19 et seq. of Regulation (EU)2019/1020 (MRS).[2]

[1] Regulation (EC) No 765/2008 of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products.

[2] Regulation (EU) No 2019/1020 of 20 June on market surveillance and product conformity.

II. Development, processing and final content

Regarding the governance structure, the proposed European Commission Regulation Title VI[3]. Although it established competences for the Commission and the Member States (MS), it focused especially on the creation of a European Committee on Artificial Intelligence (CAI), configured as an advisory, assistance and technical consultation body, which was to facilitate harmonised and consistent application, as well as cooperation with the other authorities. It would include the national AI supervisory authorities, the European Data Protection Supervisor and the European Commission itself, which would be responsible for chairing it and ensuring its operation. The multilevel scheme was essentially completed by the competent national authorities (national supervisory authority, market surveillance authority and notifying authority), which were responsible for a large part of the supervisory and enforcement functions of the Regulation.[4]

The European Council's common position (December 2022) proposed maintaining the CAI, as the body in charge of the harmonised implementation of the Regulation, and advising the Commission and MS, but broadened its tasks by requiring it to structure formulas that would permeate the positions and interests of all *stakeholders* in the ecosystem. The position strengthened the role of the Member States in the Committee, proposing the appointment of any civil servant or member of public entities, and reduced the functions of the Commission, to which it entrusted administrative and analytical support, but excluded it from participating in the Committee's votes. As for the national authorities, it established only the obligation to designate at least one market surveillance authority and one notifying authority, moving closer to the structure established by the Union's product harmonisation legislation.

For its part, the European Parliament, in amendments adopted in June 2023, proposed substantial changes. The proposal for the European AI Committee was replaced by a European Office for Artificial Intelligence, based in Brussels, with its own legal personality and the status of an independent body. While the implementation of the Regulation would remain with the

[3] A general assessment of the proposal can be found in Ebers, M., *et al.*, "The European Commission's Proposal for an Artificial Intelligence Act-A Critical Assessment by Members of the Robotics and AI Law Society (RIALS)", *J - Multidisciplinary Scientific Journal*, n.º 4, 2021 pp. 490 et seq.

[4] We have had the opportunity to comment on the governance structure of this initial proposal at another time. See Hernández Peña, J.C., *El marco jurídico de la inteligencia artificial. Principios, procedimientos y estructuras de gobernanza*, Thomson-Reuters Aranzadi, Cizur Menor, 2022, pp. 173 and ff.

national authorities, the proposal sought to strengthen the independence of the advisory and coordinating body, without radically changing its powers. With regard to the national authorities, Parliament followed a similar line to that taken with regard to the Community coordinating and advisory authority. Not only did it restore the designation of national supervisory authorities, but it also preconfigured them as independent administrations, requiring them to exercise their powers and perform their functions in an independent, impartial and objective manner, without taking instructions from any public body.

The final text departs from the Parliament's proposal and establishes, by means of a compromise solution, a more attenuated system of multilevel governance in which the MS maintain a large part of the supervisory powers, with the exception of the general purpose models whose control is communitised. Such purposes are entrusted to the AI Office, although far removed from the figure of a European Agency, as we shall see below. The Commission and the Member States retain a significant amount of competences that allow them to strategically direct the uses of AI and ensure a reliable ecosystem, through their participation in the CAI and the designation of the competent national authorities, which are assimilated to those included in the Market Surveillance Regulation. Finally, to ensure the involvement of all stakeholders and to guarantee appropriate advice, support and consultation structures are set up with various functions (Advisory Forum, expert group, standing sub-groups).

With regard to the measures and procedures for market surveillance and control of AI systems, apart from some minor modifications, the variations between the different proposals focused fundamentally on the distribution of competences in relation to the governance model that was accepted or the long-standing discussion on certain systems, which are dealt with in other chapters. Perhaps the most significant change is the establishment in the final text of a procedure for AI systems classified as non-high risk in application of Annex III.

III. The EU Governance Model of the AI Act

The European AI policy, which has its major current reference point in the AIA, aims to ensure that the use of this family of technologies is ethical[5]

[5] On this, among others, see Salazar, I., "El diseño ético de la inteligencia artificial para no discriminar ni lesionar derechos", in Balaguer Callejón, F. y Cotino Hueso, L. (Coords.), *Derecho Público de la Inteligencia Artificial*, Fundación Manuel Giménez Abad, Zaragoza, 2023, pp. 85 y

and reliable, guaranteeing an ecosystem of trust[6], reflecting the commitment to the unavoidable European values and fundamental rights.[7]

To achieve this goal, the design of a sound governance system is an essential pillar for the effective and harmonised implementation of the Regulation, and a consequent transmission belt that prohibits the trading of rights or the establishment of fragmentation spaces that allow for regulatory arbitrage. In this regard, as the now defunct High Level Expert Group on Artificial Intelligence (AI-HLEG) rightly pointed out, the development of a reliable AI ecosystem is only possible with the intervention of independent oversight mechanisms with a real capacity to expand the Union's capabilities to respond to the uncertainty introduced by the widespread permeation of AI.[8]

But before doing so, it is worth making some clarifications in order to contextualise the design of the AIA. Intervention at the Community level aims for a high level of protection of public interests, without undermining European competitiveness or compromising fundamental rights[9]. The attribution of functions and competences among the different actors of the governance system seeks to distribute responsibilities appropriately under a multilevel government paradigm, but trying to avoid excessive fragmentation by making use of a surgical application of the principle of subsidiarity, proportionality and necessity, while guaranteeing spaces for rationalisation of administrative structures and cooperation among the different actors involved.

The formula pursues a policy cycle approach that allows for systematic

ss.; Moreno Rebato, M., *Inteligencia artificial (Inteligencia artificial (Umbrales éticos, Derecho y administraciones públicas)*, Thomson Reuters Aranzadi, Cizur Menor, 2021, *in totum*; Cotino Hueso, L., "Ética en el diseño para el desarrollo de una inteligencia artificial, robótica y big data confiables y su utilidad desde el Derecho", *Revista Catalana de Dret Públic*, n.º 58, 2019, pp. 29 et seq.

[6] Gamero Casado, E., "El enfoque europeo de inteligencia artificial", *Revista de Derecho Administrativo - CDA*, n.º 20, 2021.

[7] The doctrine has already pronounced itself on the meaning and ultimate aims of the Regulation. In this regard, see Cotino Hueso, L., "Un análisis crítico constructivo de la Propuesta de Reglamento de la Unión Europea por el que se establecen normas armonizadas sobre la Inteligencia Artificial (Artificial Intelligence Act)", *Diario La Ley*, 2021.

[8] AI HLEG, *Policy and Investment Recommendations for Trustworthy AI*, 2019, p. 37.

[9] On the impact of AI on fundamental rights, see the interesting studies by Cotino Hueso, L., "Nuevo paradigma en las garantías de los derechos fundamentales y una nueva protección de datos frente al impacto social y colectivo de la inteligencia artificial" in Cotino Hueso, L. (Dir.), *Derechos y garantías ante la inteligencia artificial y las decisiones automatizadas*, Aranzadi, Cizur Menor, 2022, pp. 69 ff; Simón Castellanos, P., "Taxonomía de las garantías jurídicas en el empleo de los sistemas de inteligencia artificial", *Revista de Derecho político*, n.º 117, 2023, p. 155 et seq.; and Aba Catoira, A., "La garantía de los derechos como respuesta frente a los retos tecnológicos" in Balaguer Callejón, F. and Cotino Hueso, L. (coords.) *Derecho Público de la Inteligencia Artificial*, Fundación Manuel Giménez Abad, Zaragoza, 2023, pp. 57 et seq.

monitoring, from the development and evaluation of a complex policy programme, to the ongoing evaluation of AI systems throughout their life cycle. This model structures the regulatory framework and governance entities concentrically. The implementation and enforcement of the regulatory cascade that has the Regulation as its peak, but which includes technical and ethical standards guiding the design, modelling, commissioning, and decommissioning of AI systems by providers, represents the first circle of protection. The implementation of this regulatory cascade will be verified by the Notified Bodies, discussed in other chapters. In turn, these notified bodies have to be previously accredited by the notifying organisms, which are responsible for verifying that they have the technical requirements to carry out their work, and to reinforce the standard of protection required at Community level. This constitutes the second circle. Finally, once they have been placed on the market, it must be ensured that the systems in operation comply with the regulations and do not entail risks, with a significant amount of powers for *ex-post* control and supervision, through monitoring and control procedures. These functions are attributed to the market surveillance authority or the AI Office, depending on the characteristics of the particular AI system.

As mentioned above, we will first deal with the governance authorities and bodies, and later with the monitoring and control procedures set out in the regulation itself, as well as in the general framework of the new Community harmonisation approach.[10]

IV. European Commission. Powers and functions

Although the AIA includes a small galaxy of bodies, administrative, and support structures for its implementation, it gives the Commission an important role in ensuring its harmonious and coherent application. Given that the structure and functioning of this Community institution is well known, we will now refer to the important group of regulatory, supervisory and control powers reserved to it under the AIA.

The Commission plays a key role in the regulatory environment. It is responsible for contributing to the final design of the regulatory programme, as well as ensuring that it is adapted to technological developments. To this end, Art. 97 AIA empowers it to amend non-essential issues by adopting del-

[10] Álvarez-García, V. and Tahirí Moreno, J., "La regulación de la inteligencia artificial en Europa a través de la técnica armonizadora del nuevo enfoque", *Revista General de Derecho administrativo*, n.º 63, 2023, p. 5.

egated acts[11], while Art. 175 and the articles of the text empower it to adopt implementing acts[12] to ensure its uniform application.

However, on the basis of delegated acts, it is entrusted, inter alia, to establish the methodology and list of criteria for classifying high-risk stand-alone systems (Recital 52); with regard to general purpose models, to specify and adjust the criteria and indicators for comparison, including the threshold for classifying them as systemic risk in accordance with Annex XIII, and to designate them (Art. 52.4); or to amend the minimum elements of the technical documentation to be completed by the providers of general purpose models in Annex XI and the transparency obligations established by Annex XII (Recital 101 and Art. 53.5). and Art. 53.5).[13]

Furthermore, by means of implementing acts, it is empowered to approve codes of good practice for the fulfilment of obligations by providers of general-purpose models (art. 56.6). Also rules regulating the labelling and detection of manipulated or artificially generated content, with the assistance of the AI Office (art. 50.7); defining requirements and modalities for the creation, operation and supervision of controlled test areas (art. 58), as well as the testing of high-risk systems in real time (art. 60.1).[14]

In addition to the powers to issue delegated and implementing acts, it is

[11] In line with Art. 290 TFEU, in full respect of the principles of proportionality and subsidiarity, following the guidelines of the Interinstitutional *Agreement Between the European Parliament, The Council of The European Union and The European Commission on Better Law-Making*, 13 April 2016. On this, Bradley, K. S. C., "Legislating in the European Union", in Barnard, C. and Peers, S., *European Union Law*, second edition, Oxford University Press, New York, 2017, pp. 126 et seq.

[12] In accordance with Art. 291 TFEU. These powers, suffice it to say, are to be exercised with respect to the principle of proportionality and subsidiarity, as well as Regulation (EU) 182/2011 of the European Parliament and of the Council of 16 February 2011 laying down the rules and general principles concerning mechanisms for control by Member States of the Commission's exercise of implementing powers.

[13] As might be expected, the list of delegated acts does not end with those mentioned above. Without being exhaustive, it is also responsible for amending the list of Union harmonisation legislation for both the old and new approaches (contained in Annex I); adapting the conformity assessment procedures; the content of the EU declaration of conformity contained in Annex V; and the regulation of conformity assessment procedures based on internal control mechanisms and assessment of quality systems in Annexes VI and VII, in order to ensure that they are effective (rec. 173 and Art. 47.5).

[14] In addition to the above, it is empowered by implementing acts to set up the Independent Scientific Expert Group (Art. 68.1); to regulate the participation of independent experts in the assessments of general purpose models (Art. 92.6); or to challenge the notification of a non-compliant notified body, with the power to suspend or withdraw the notification if the notifying MS fails to take appropriate corrective action (Art. 37.4).

empowered to issue *soft law*. It is empowered to promote guidelines related to Annex III systems that are not considered high risk by exception[15], or the adoption of methodologies, measurement and benchmarking indicators to ensure an adequate level of accuracy, robustness and cybersecurity (Art. 15.2). It can also issue guidelines providing guidance on the fundamental elements of simplified quality management systems for high-risk AI systems (Art. 63.1) applicable to SMEs, in order to calibrate compliance costs, and provided that they do not result in lowering the standard of protection of fundamental rights or other overriding purposes of general interest.

In addition to the above-mentioned responsibilities, the Commission is entrusted with supervisory and control functions. They reach a high intensity with regard to the compliance of general purpose model providers with their obligations, the supervision of which is communitarised (Art. 88)[16]. However, this function is to be delegated to the AI Office, and we will come back to this later.

Whether it is for the Commission to issue individual decisions designating general purpose systemic risk models[17], as well as to reassess or rate them on the basis of qualified alerts or otherwise (Art. 51 and 52). For these purposes, it may request from the provider the information to assess the system (Art. 91), including access to the source code through application programming interfaces (APIs) or other technical means (Art. 68.3). It is also competent to impose fines on the providers of these models for the infringements listed in Art. 101.1.

Supervisory powers extend to the exercise of powers by MS. The aim is to ensure a high degree of harmonisation, especially with regard to the areas that confer deference to the MS in order to complete the regulatory programme. It is important to identify divergences that create areas of fragmentation of the common market or tend to lower the standard of protection of fundamental rights. Real-time biometric identification systems, discussed in another chapter, are an example of areas that may lead to pathological practices. Hence, the Commission is empowered to collect information on national rules adopted by Member States (MS) to exceptionally authorise their

[15] It is true that this qualification will be carried out by applying the criteria set out in the Regulation itself (art. 6.3), but the Commission may adopt guidelines specifying how they are to be applied, as well as a list to help identify them (art. 6.5).

[16] This approach is intended to maximise and centralise the ability to bring together expertise and know-how, ensuring consistent implementation and oversight to respond to the risks introduced by these systems.

[17] On the basis of the criteria set out in Annex XIII or equivalent.

use (5.4), and market surveillance and data protection authorities are required to submit an annual report on the use of such systems (Art. 5.6).

Moreover, in order to ensure compliance with transparency obligations and to promote the acceptance of AI, the Commission must establish and maintain a database in which high-risk system providers other than those subject to harmonised legislation, as well as those in Annex III that by way of exception are not considered as such, are registered (Art. 71).

Finally, to close the block of powers aimed at ensuring harmonised application, it is up to the Commission to evaluate and assess the Regulation, making proposals for updating it to Parliament and the Council (Art. 112).

The above competences relate primarily to the regulatory programme or the exercise of legal powers. However, the Commission is also entrusted with functions related to other technical means that also aim to promote confidence in the use of AI. Thus, on the one hand, it has a leadership role to play in AI literacy[18], in order to promote its acceptance by citizens[19]. To achieve this objective, art. 4 sets out a series of obligations, on the understanding that the development of these capacities is crucial for the AI ecosystem value chain to be able to adequately comply with the regulations. To this end, with the support of the AI Committee[20], it is responsible for developing and promoting literacy tools that enable users and affected parties to understand the risks and benefits, as well as the rights and obligations deriving from the Regulation and the entire regulatory framework that develops it. It is also empowered to develop, in collaboration with MS and the AI Office[21], voluntary codes of conduct for developers, deployers or users (cons. 20) for similar purposes.

Also with regard to powers related to other non-technical instruments, the Commission is responsible for promoting the adoption of technical stan-

[18] This is understood as the development of skills, knowledge and understanding that will enable providers, users and stakeholders, taking into account their rights and obligations, to make an informed deployment of AI systems, as well as to become aware of the opportunities and risks or possible harms that this technology may cause (cons. 56), including in our understanding its social, ethical and environmental impact.

[19] This attribution responds to the commitment made by MS and the EU in different instruments. The Ethical Guidelines for Trustworthy AI list it as one of the key non-technical mechanisms for generating a trustworthy AI ecosystem, complementing regulatory response, codes of conduct or standardisation. AI HLEG, *Ethics Guidelines for Trustworthy AI*, pp. 22-23. In the same vein, the UNESCO Recommendation on the Ethics of Artificial Intelligence, approved with the support of the EU MS, includes it among its principles, as well as incorporating it among the public policies they are committed to implementing. UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, 2021, pp. 23 and 33.

[20] Art. 66.f AIA assigns this function.

[21] In this sense, art. 95.2.c AIA.

dards that reflect the state of the art and contribute to maintaining a high level of protection for citizens.[22]

Although the final development of these technical standards is the responsibility of the European standardisation institutions (CEN, CENELEN and ETSI), it is up to the Commission to adopt the mandate and entrust them with their development (Art. 40.2). Moreover, in the absence or inadequacy of these standards, it is empowered to issue common specifications (Art. 41), which means that it must ultimately ensure the existence of instruments - even exceptional ones[23] - that allow providers to demonstrate their conformity, encouraging innovation without reducing the protection of citizens.

V. AI Office. Nature, structure and functions

The creation of the AI Office (AIO) was discussed in the trilogues. The Commission's initial proposal did not foresee it, while the Parliament's mandate established it by absorbing part of the functions of the AI Committee. In the end, the agreement of the intergovernmental negotiations established a brief mandate to the Commission for its establishment.

The AIO is now established by Art. 64, although its functions are differentiated from those attributed to the AI Committee. The Commission implemented the mandate, prior to the adoption of the final text, by Decision C(2024)390 of January 2024 (hereinafter the Decision), and it became operational on 21 February of the same year.

The Regulation is brief with regard to its nature, structure or integration into the Commission, and it is the Decision that provides us with these coordinates. Although at the Community level the distinction between specialised offices and agencies can be somewhat blurred, in our case the regulation creating the AIO provides some clarity. We are not dealing with an agency, and so we are moving away from the techniques and characteristics of these Community administrative bodies typical of the European composite administration[24]. Nor is it an entity that can be recognised as having its own legal personality, as is the case with specialised agencies.

In fact, AIO departs from the European entities within the meaning of

[22] On the meaning and functioning of this type of rules, see Álvarez García, V., *Derecho de la regulación económica. VIII. Industria*, Iustel, Madrid, 2010; Álvarez García, V., *Las normas técnicas armonizadas (una peculiar fuente del Derecho europeo)*, Iustel, Madrid, 2020.

[23] See in this respect recital 121.

[24] On this issue, among others, Schmidt-Assmann, E., "European administration by European agencies", *Lex Social*, Vol. 3, No. 2, 2013, pp. 5 et seq.

the Financial Regulation 2018/1026[25]. It is constituted as an integral part of the administrative structure of the Directorate-General for Communication Networks, Content and Technology (DG CONNECT), and is subject to its annual management plan[26]. For this reason, it has to operate in accordance with the Commission's internal procedures (Recital 7 of the Decision), maintaining close coordination with MS, competent national authorities, as well as other bodies, agencies and specialised support instruments, such as the Scientific panel of independent experts that will support it in the development of its activities (art. 68.3). In short, it is an integrated structure within the Commission with a certain operational autonomy. [27]This accentuates an idea expressed previously. The institutional design of the Regulation reserves to the Commission an intense supervision and strategic control of the AI, in order to integrate it cohesively with the rest of the Community policies of the single digital market.

This applies in particular to its financial and human resources, although it enjoys a certain degree of autonomy in that it has its own staff chapter. Its staff is composed of members already assigned or reassigned to DG CONNECT, although Article 8 of the Decision allows for external staff to be recruited at the cost of redistribution of budget lines from the Digital Europe Programme[28]. The same instrument will cover its operational expenditure, in line with the specific objective set out in art. 5 of Regulation 2018/1046. This ensures that the Office is provided with resources and qualified staff, with the necessary expertise, but without unduly burdening the Commission's budget.

The Articles of the AIA, as well as Art. 3 of the Decision, set out the powers and functions of the AIO. Their main purpose is to ensure that the Commission and the other governance structures have the capacities and expertise to ensure the compliance and risk mitigation of general purpose models. For the purposes of systematisation, these powers will be grouped into three distinct blocks, although they should be interpreted in an integrative perspective.

[25] Regulation (EU, Euratom) 2018/1046 of the European Parliament and of the Council of 18 July 2018. Article 2.26 of this regulation defines European offices as "an administrative structure set up by the Commission, or by the Commission together with one or more other institutions of the Union, to perform specific cross-cutting tasks".

[26] Cons. 6 and Art. 1 of Decision C(2024)390 final.

[27] This integration is clearly stated in Art. 3.47 by defining AI as "the Commission's role in contributing to the implementation, monitoring and supervision of AI systems and to the governance of AI".

[28] Funding instrument covered by Regulation (EU) 2021/694 of the European Parliament and of the Council of 29 April 2021 establishing the Digital Europe Programme.

First, there is a large group of functions and competences related to general purpose models, to which we will pay more attention. For these, AIO acts as a transmission belt of expertise to ensure that the Commission and the other authorities respond proactively and reactively to risks to the functioning, explainability and transparency of these models, to assess them and to determine whether the regulatory programme is adequately applied to them. In this regard, it is responsible for developing the tools, methodologies and benchmarks for their assessment, especially for those models that present systemic risks (Art. 3.1.a of the Decision).[29]

It is also responsible for the supervision at EU level of general purpose models, an attribution of particular significance given the concerns that have been generated by the regulation of these models during the negotiation of the Regulation. In this regard, it should be recalled that the Commission is granted exclusive powers to supervise compliance with the obligations of providers (Art. 162 and Art. 88.1 AIA); a function which it must entrust to the AIO, as required by Art. 88.1. This task is of varying intensity, particularly in the case of systems that are based on general purpose models and both model and system are supplied by the same provider. In this case, it assumes their surveillance and control, and also becomes the Market Surveillance Authority (MSA), with the competences and functions of these bodies (art. 75).[30]

With respect to other general purpose systems, Art. 89 empowers AIO to take follow-up actions and measures to ensure proper implementation by providers[31]. In line with the above, it is also empowered to conduct investigations into possible breaches by providers of general purpose AI models and

[29] The formulation of these instruments is crucial to understand the potentialities and limitations inherent in these models and, consequently, to surgically modulate the intervention of the actors in the governance chain. Hence the particular importance of reviewing and updating the methodologies and classification thresholds of general purpose models with systemic risks.

[30] In the exercise of these powers, the Office may collect complaints and claims from any person or entity that has grounds to believe that infringements have been committed (art. 85). The same is recognised with regard to intermediate providers, who are empowered to file complaints for justified infringements and in compliance with the provisions of Art. 89.2, as well as professional representatives who terminate their mandate in the event of non-compliance by the supplier with the obligations set out in the AIA (Art. 54.4).

[31] Thus, after consulting the AI Committee, it is responsible for assessing whether the information provided to the Commission is insufficient or whether it is appropriate to investigate possible systemic risks on the basis of qualified reports from the Independent Scientific Expert Group, with activation of the alert system (Art. 90.2), and it is empowered to establish structured dialogues with providers (Art. 91.2).2); also to receive information and notifications

systems, including non-compliance breaches, both on its own initiative acting as an AVM and at the request of national AVMs.[32]

Finally, continuing in this first block of powers, it is empowered to promote and collaborate in the approval of codes of good practice at EU level aimed at general purpose models with systemic risk[33]. A similar provision is included with respect to obligations relating to the detection and labelling of artificially generated or manipulated content, in which case it is responsible for drawing up codes of good practice (art. 50.7).

A second set of responsibilities empowers the Office to promote effective coordination between administrative bodies in the governance ecosystem by facilitating their assistance or exchange of information. In this regard, it is tasked with supporting the implementation of rules on prohibited practices, as well as high-risk systems, with a view to achieving a high degree of harmonisation and cohesion so as not to fragment the single market or allow regulatory arbitrage by operators. It is also given a number of functions related to the implementation of controlled testing areas[34]; in the area of market surveillance actions[35]; and the development of codes, criteria, clauses and templates to facilitate the application of the Regulation by national authorities and other *stakeholders*, helping to manage the complexity of its implementation and enforcement.[36]

from providers in the event of activation of the classification procedure of a general purpose model with systemic risks, as well as of serious incidents related to these models and of the corrective measures adopted to mitigate their impact (art. 55.1.c); and to request the technical documentation of models that are not free and open (art. 53.1.a and 54.2), as well as the information necessary to demonstrate compliance with the obligations set out in Chapter V. With regard to the latter, it is worth recalling that models made available under a free and open licence are exempt from this obligation, unless they are considered to be of systemic risk (art. 53.2).

[32] Thus, Art. 92, in line with recitals 163 and 164.

[33] In cooperation with the AI Committee, it is responsible for assessing providers' adherence, identifying non-compliance or inconsistencies in its implementation and publishing reports on the achievement of its objectives (art. 56).

[34] While the implementation and development of *sandboxes* is the responsibility of the AIO (Art. 57.1), those authorities must be informed and are empowered to provide advice and guidance on request (Art. 57.15). In addition, these authorities must report on temporary or permanent suspensions of trials, and submit an annual report on the operation of sandboxes to the AIO (Art. 57.11 and 16).

[35] The AIO is responsible for supporting and coordinating the conduct of joint investigations (Art. 74(11)). It is also responsible for providing support in the case of general purpose model investigations by national market surveillance authorities, using the cross-border mutual assistance procedure provided for in Art. 22 et seq. of Regulation 2019/1020.

[36] Thus, together with MSs, it is responsible for facilitating the development of volun-

The last block contains powers related to assisting the Commission in the preparation of delegated and implementing acts and decisions, in order to ensure consistent application of the Regulation. This includes assisting the Commission, including urging, preparing and updating guidelines (Art. 96.2), as well as developing a methodology and guiding revisions of the criteria for assessing risk levels, and assessing the inclusion of new systems in Annex III, in the list of prohibited practices and of systems requiring additional transparency measures (Art. 112.11).

VI. Competences of Member States and competent national authorities

The AIA assigns a key role to MS in its governance architecture. In addition to designating national supervisory authorities, which will be discussed below, they participate in Community governance bodies such as the CAI, designating a representative to act as a single point of contact (Art. 65.2). They are also given the power to develop the system of sanctions and enforcement measures and to implement it in accordance with the Commission's guidelines (Art. 96 and 99).

Moreover, MS are key players in ensuring the modality of the market surveillance governance network[37]. In this respect, it should be recalled that the Regulation, once adopted, became part of the harmonised product safety legislation[38], so cooperation between the different authorities in charge of

tary codes of conduct for system providers that are not considered high-risk (Art. 95). It is also responsible for developing voluntary model contractual clauses for providers of high-risk systems and third parties using or integrating such systems (25.4); providing templates for: (i) collecting by providers of general purpose models a detailed public summary of the data used for training (Art. 53.1.d); (ii) develop and maintain unified information on Union operators; (iii) raise awareness of the obligations established by the Regulation through communication campaigns; and, (iv) promote convergence of best practices in the area of public procurement of AI systems (Art. 62.3.d). It is also responsible for preparing an automated questionnaire for fundamental rights impact assessments of high-risk systems for those responsible for the deployment of these systems, in order to facilitate the assessment of their compliance with the obligations of Art. 27. On these assessments in general, see Simón Castellanos, P., *La evaluación de impacto algorítmico en los derechos fundamentales*, Aranzadi, Cizur Menor, 2023.

[37] On modality as a mechanism of digital governance, see the classic work by Hood, C. and Margetts, H., *The Tools of Government in the Digital Age*, Palgrave, Hampshire, 2007, pp. 21 ff.

[38] This can be concluded from the reference and application of Regulation (EU) 2019/1020 which, as pointed out in the Blue Guide on the application of European product legislation (2022/C 247/01), is one of the formulas to delimit its scope beyond what is foreseen by its Art. 2 and Annex I of the MRS. It is also clear from the structure of the AIA that it takes over the reference provisions of the Community product harmonisation legislation

these functions must be ensured. This is confirmed by the AIA[39]. They also have an important role to play in ensuring the application of the principle of innovation, since it is ultimately up to them to ensure that the designated authorities set up and operate at least one *sandbox*, or to guarantee this obligation by equivalent national coverage, as explained elsewhere in this work.

Finally, they have some implicit competences. Thus, in application of art. 13 of Regulation (EU) 2019/1020, they must consider and incorporate into the national market surveillance strategy the priorities for ensuring and effectively supervising compliance with the AIA. Likewise, and given that defence and national security are exclusive competences of MS, it will be up to them to determine and regulate the AI systems intended for these purposes, in strict compliance with the regulatory perimeter of exclusion established in Recital 24 and art. 2.3.

In addition to the above, MS complete the governance architecture at the national level, as they designate the competent national authorities in charge of the supervision and enforcement of the Regulation within their jurisdictions. In relation to this, the AIA replicates the structure foreseen in the product harmonisation regulation[40], by providing for the obligation to designate at least one notifying authority and one market surveillance authority (Art. 70.1)[41]. The concentration of all these authorities into a single authority is allowed, although it will assume all the tasks attributed to it. With this deference, States are empowered to choose the formula they consider most appropriate according to their internal organisation, respecting the mandate of Art. 5 TEU. They may resort to a system of functional or geographical distribution of powers[42], provided that the uniform application and effectiveness of the Regulation is guaranteed.

laid down by Annex I to Decision No 768/2008/EC of the European Parliament and of the Council of 9 July 2008 on a common framework for the marketing of products. However, for the avoidance of doubt, recital 76 of the AIA expressly states this.

[39] In this regard, they are to facilitate cooperation with the competent authorities in application of the harmonisation rules in Annex II or the high-risk systems in Annex III (Art. 74.10). It also extends to the responsibilities assigned to the AIO, the review tasks of the Commission mentioned in Article 112 and the national authorities in charge of the protection of fundamental rights, such as the Spanish Data Protection Agency.

[40] Specifically Regulations (EC) 765/2008 and (EU) 2019/1020.

[41] This is in line with Art. 10.1 of Regulation 2019/1020, which attributes exclusive competence to MS to assign these authorities and to supervise the market within their territory.

[42] A deference that facilitates the distribution of competences between the different levels of government in federal or decentralised states, by allowing the distribution of supervision between authorities with different internal territorial competences. This is, for example, the case in Spain, where market surveillance competences are attributed to the Autonomous Com-

Whether one or more national authorities are designated, it must be ensured that they act impartially and objectively. Their autonomy should also be guaranteed by providing them with adequate permanent financial and human resources for the exercise of their functions.[43]

In addition, the final version of the AIA requires such authorities to be independent and impermeable to any activity incompatible with their functions. Although it is not the place to pronounce at length on this issue, in our view the terms of the Regulation and Recital 154 are not strong on the requirement to establish or designate an independent authority[44]. Despite this, in the case of Spain, some of the administrations included in Law 40/2015, of 1 October, on the Legal Regime of the Public Sector (LRJSP), may not be the most suitable for maximising the mandate of independence required by the AIA. In this regard, the formula adopted for the creation of the Spanish Agency for the Supervision of Artificial Intelligence (AESIA), despite meeting the requirements of the Regulation, may be deficient in terms of the number of members with technical expertise and a desirable reinforcement of its independence.[45]

The designation or establishment of national authorities must be communicated to the Commission to facilitate collaboration and joint action with the other competent national authorities, and to ensure that they comply with and have the necessary powers to carry out their supervisory functions. The Commission is empowered to request information, and MS are obliged to

munities. This issue is analysed at length by Izquierdo Carrasco, M., *La seguridad de los productos industriales. Régimen jurídico-administrativo y protección de consumidores*, Marcial Pons, Madrid, 2000, pp. 65 ff.

[43] For these purposes, experience and expertise in Artificial Intelligence and data science technologies, as well as in legal issues that allow for the assessment of risks to fundamental rights, security or health (Art. 70.3), must be taken into consideration. Furthermore, compliance with confidentiality obligations in their actions must be ensured (Art. 78).

[44] In line with the above, the Regulation leaves Member States free to designate any authority as long as the above-mentioned conditions are met. Recital 153 expressly states that "Member States may decide to designate any type of public entity to carry out the tasks of national competent authorities within the meaning of this Regulation, in accordance with their specific national characteristics and organisational needs".

[45] It should be noted that AESIA was set up as a State Agency under Articles 108 bis et seq. of the LRJSP. Although the Statute of the Agency, approved by Royal Decree 729/2003, recognises a certain degree of independence of a technical nature, it does not include guarantees of independence for the exercise of the functions of its members. In fact, the Presidency of AESIA is held by the Secretary of State for Digitalisation and AI, who also holds the Presidency of its Governing Council, the majority of whose members are senior officials of the General State Administration. To remedy these shortcomings, the ideal solution would be to approve by law a truly independent AI authority, in accordance with art. 110 LRJSP.

submit a report for discussion and recommendations. This is intended to promote harmonised action, as well as incremental co-ordination through indicator-based governance arrangements[46] and peer review[47] widespread in the area of product safety.[48]

However, regardless of the institutional structure of each MS, some general powers are entrusted to the competent authorities. Thus, among others, they are responsible for providing guidance and advice on the application of the AIA, especially to small-scale providers and *start-ups* (Art. 70.8). Also, with regard to high-risk systems, they are empowered to supervise compliance with horizontal requirements; the regulatory adjustment of those that are not considered high-risk by exception; and they are empowered to require providers and other actors to comply with information and communication obligations.

That said, the powers and responsibilities of market surveillance authorities and notifying authorities will be examined concisely below, although it should be noted that many of these are explained at length in other chapters of this paper.

1. Market surveillance authorities

The market surveillance authority (MSA) is defined as the national authority responsible for the surveillance functions and the adoption of measures provided for in Regulation 2019/1020[49]. It should be recalled that market surveillance is structured at national level in accordance with the provisions of this Regulation. Therefore, in addition to the above functions, the MSA must be empowered to carry out coordination and cooperation tasks at Community level, as required by art. 10.4 MRS. This is why the Regulation designates it as the single point of contact with the Commission and the other authorities (Art. 70.2 AIA).

In relation to the latter, the articles of the AIA attribute to it a set of

[46] On governance by indicators see, among others, Davis, K., Kingsbury, B. and Merry, S., “Indicators as a Technology of Global Governance”, *IILJ Working Paper*, no. 2010/2, 2010.

[47] On *Peer Review and Peer Pressure as a* governance technique, see Baldwin, R., *et al.*, *Understanding Regulation. Theory, Strategy, and Practice*, second edition, Oxford University Press, New York, 2011, p. 431.

[48] See, for example, Art. 12 of Regulation 2019/1020.

[49] It should be noted that the exception is the European Data Protection Supervisor, which according to Art. 74.9 will act as a market surveillance authority when systems falling within the scope of the AIA are deployed by EU institutions, bodies, offices and agencies, with the exception of the CJEU in its jurisdictional role.

powers and obligations, which are complemented by those set out in Article 14 of the MRS. Without being exhaustive, it must report to the Commission, on a regular basis, on activities related to market surveillance and on information of potential interest related to competition law[50]. Moreover, and given that market surveillance is aimed at protecting consumers, it is entrusted with informing and raising awareness among citizens and operators, guaranteeing the principle of transparency and the mandates of the national regulations that develop them.[51]

In terms of market surveillance functions, they are competent to supervise and control the safety requirements of the Regulation. In particular, they are given competences to supervise AI systems designed as safety components of other products subject to the "new approach" regulation, giving them intense powers[52]. Finally, they are given competences related to some conformity assessment procedures, as well as to post-market surveillance procedures and measures, which have already been studied and we refer to the corresponding sections.

2. Notifying authority

The notifying authority is the body responsible for designating and notifying conformity assessment bodies at Union level, in accordance with harmonised legislation and the AIA. In this regard, it is responsible for establishing and implementing procedures related to the assessment, designation, notification and monitoring of conformity assessment bodies (Art. 3.19). Essentially, they are responsible for assessing the ability and technical competence of notified bodies to implement the conformity assessment procedures of AI systems, in a full, impartial and independent manner, and for reporting their designation to the Commission and other Member States through the corresponding systems.[53]

[50] In this case, Art. 74(2) further requires that the national competition authorities be notified.

[51] In the case of Spain, Law 19/2013 of 9 December on transparency, access to public information and good governance.

[52] Thus, to highlight just a few, they can request access to data of various types, including data used to train and validate models, even using technical means that allow remote access, such as APIs and programming interfaces. They can also, upon request and with reasonable justification, access the source code of high-risk systems in order to assess compliance with horizontal requirements.

[53] This is through the NANDO (New Approach Notified and Designated Organisations) information system, which is administered by the Commission.

Although these bodies and their notification procedures are explained in detail in other chapters of this work, it is important to note that the role of the notifying authority goes beyond the designation mentioned above, as it ultimately assumes responsibility for the technical competence and capability of the bodies it notifies. Moreover, although it is the responsibility of the notifying authorities to establish the designation and notification procedures, both the general product safety regulation[54] and the Regulation provide guidelines to guide this work. They should be proportionate, avoid unnecessary burdens on providers in application of the innovation principle, and take into account circumstances such as the size of companies and the specific AI system.

In terms of their performance, the Regulation sets out a number of principles that reflect the mandates of the general regulation[55]. Thus, they must be impartial and objective, avoid conflicts of interest with respect to both the activities and the participants in the evaluation processes, and guarantee personal and technical competence in the exercise of their functions. Their regime, as is necessary, is completed with obligations of confidentiality and secrecy regarding the information to which they have access; the prohibition of profit-making; as well as the principle of non-competition with notified bodies.

VII. European Committee on Artificial Intelligence and the permanent sub-groups

The European Committee on Artificial Intelligence (CAI) is another key part of the institutional architecture created by the AIA, as its remit is not only to contribute to its harmonious implementation, but also to reflect the diversity of interests in the EU AI eco-system (Recital 149). In addition, a number of permanent sub-groups are set up within the AIA to deal with specific issues. These are discussed below.

1. European Committee on Artificial Intelligence. Structure and terms of reference

The membership of the CAI is mainly composed of MS representatives. The European Data Protection Supervisor also participates, but as an ob-

[54] In particular Decision 768/2008/EC, cited above.

[55] Art. 4 of Regulation (EC) 765/2008. In domestic legislation, see Article 17 of Law 21/1992 of 16 July 1992 on Industry.

server. The initial versions of the proposed Regulation provided for a more active participation of the Commission, which was to chair it and manage its meetings. However, the final version provides for it to be represented by the AIO, although without voting rights.

In addition to these natural members, other national authorities, bodies, or national or Union experts may participate -upon invitation- if the issues to be addressed are of relevance to them. It is striking that the AIA radically excludes the participation of international experts in view of the advisory role to be played by the Committee. This limitation potentially restricts access to knowledge on sophisticated developments, and deprives regulators of valuable experience and perspectives for the proper ethical and legal regulation of AI.[56]

Turning to the MS representatives, they will be appointed for a period of 3 years, once renewable. In contrast to the initial proposal, which assigned representation to national supervisory authorities, the AIA makes it explicit that States may flexibly designate persons linked to any public body, provided that they have the competence to coordinate implementation internally and have the power to contribute to the development of the Committee's functions.[57]

The internal functioning of the CAI will be approved by the representatives of the MS. In addition, the AIA removes the Commission's powers and protects its Presidency, which will be held by one of these representatives elected by means of the agreed procedure and formulas. However, the Statute approved in exercise of these powers of self-organisation must, in any case, guarantee that the Committee's actions are objective and impartial. It is up to the Commission to provide the administrative structure for its operation

[56] Raising the Regulation to a global standard, as the EU intends, should not be achieved through extraterritorial application. On this issue, see, among others, López-Tarruella Martínez, A., "El reglamento de Inteligencia Artificial y las relaciones con terceros Estados", *Revista electrónica de estudios internacionales*, n.º 45, 2023. The permeation of the Brussels effect is also associated with the global recognition and legitimacy of EU regulations. Therefore, the participation of experts - on issues deemed relevant and in a considered manner - would allow for the inclusion of concerns or interests worthy of protection that may be off the radar of European actors, and contribute to the acceptance of AIA as the "*Gold Standard*" of AI.

[57] In the case of Spain, this representation could be exercised by the Secretary of State for Digitalisation and Artificial Intelligence, either through the head of this Secretariat or the Directorate General for Digitalisation and Artificial Intelligence, given the broad powers of representation before European institutions set out in paragraph d) of Article 8.1 of Royal Decree 403/2020 of 25 February, which develops the basic organic structure of the Ministry of Economic Affairs and Digital Transformation, as well as the recognition of coordination and cooperation powers both at inter-ministerial level and with other public administrations.

through the AIO, offering specialised administrative and technical support, so that the proposals and recommendations are sound and based on objective elements.

The CAI is essentially a body for technical advice, assistance and consultation of the Commission and MS. It is attributed an important set of powers and functions by Art. 66, the ultimate aim of which is to facilitate the harmonised and consistent application of the Regulation, as well as cooperation with market surveillance authorities. By bringing together national and EU authorities, it is called upon to provide guidance on emerging AI issues affecting the single market, as well as to collect and make available best practices and expertise to MS and the Commission.

Within the framework of its advisory activity, it is responsible for issuing opinions, recommendations and guidelines, as well as reports on the implementation of the Regulation[58]. On the other hand, and given the variable regulatory density of the Regulation, in the complex distribution of competences established, it is up to the Committee – at least partially – to contribute to completing and revising the regulatory programme. In this regard, it is responsible for preparing delegated and implementing acts, as well as collaborating with the Commission in the periodic reviews of the Regulation provided for in Article 112. It is also responsible for approving *soft law* rules in order to standardise the administrative practices of the MS[59]. In addition, the articles of the Regulation also attribute to it functions related to advice[60], and even some supervisory and control functions in co-responsibility with other actors in the governance structure.[61]

Finally, following the tiered approach and the multi-level distribution of

[58] In this respect, it is responsible for giving its opinion on the technical and organisational capacities of MS; for reporting and proposing recommendations on harmonised standards and, where appropriate, common specifications for high-risk systems; and, also for these, for assessing possible amendments to the list contained in Annex III.

[59] In this respect, reference is made to two areas. Firstly, those referring to issues where MS must cooperate intensively to ensure harmonised implementation, such as conformity assessment procedures or innovation support measures. Secondly, those aimed at interpreting technical or legal concepts and knowledge in a uniform manner by authorities and operators, whereby the Committee should provide guidelines or benchmarks.

[60] Thus, the Commission is mandated to submit proposals for recommendations to the Commission related to the annual reports of MS on the adequacy, financial capacity and human resources of the competent national authorities (Art. 70.6); or to encourage and facilitate, together with the Commission, the elaboration and adoption of voluntary codes of conduct aimed at the fulfilment of environmental requirements (Art. 112.7).

[61] In this regard, in collaboration with AIO, it is responsible for assessing compliance with the objectives set out in the codes of conduct referred to in art. 112.7.

responsibilities for oversight and control of Generative AI, the CAI should formulate recommendations and strategic guidance regarding the implementation of the AI Act. The intensity of its involvement will vary according to the two-tier categorisation established for this type of system. Thus, it will provide guidance and advice for general purpose systems, but for qualified systems or those with systemic risks, it will take additional actions. In such cases, it will be consulted by the AIO on the need for specific general purpose model assessments; and it will provide opinions to the Commission on qualified warnings (art. 90 and 92).

It remains to be noted that the complex framework of competences outlined above does not fully shape the mechanisms through which the Committee will promote coordination, exchange of information, and cooperation between the actors in the governance structure, nor the instruments that will allow the interests of the different actors in the AI value chain to be permeated and assessed. This is because these attributions are materialised through the Terms of Reference introduced by the Council, establishing the obligation to create stable sub-groups, as mentioned above, which we will now examine.

2. Standing sub-groups for market surveillance and notifying authorities

The aforementioned Art. 65.6 of the AIA empowers the Committee to set up temporary or permanent groups or sub-groups to deal with specific issues of the Regulation. However, it mandates the creation of the Standing Sub-Group on Market Surveillance (SSMS) and the Standing Sub-Group on Notifying Authorities (SSNA).

The SSMS, which acts as a tool for exchange between market surveillance authorities, is associated with the stable cooperation mechanism that was created in application of Art. 29 of Regulation (EU) 2019/1020[62], i.e. the EU Product Conformity Network (EUPCN)[63]. As can be seen, this is another manifestation of the configuration of the AIA as one more piece of the dense network of EU product harmonisation legislation.

The incorporation of the Sub-Group into the EUCPN is intended to streamline market surveillance practices in the EU and to reinforce the effec-

[62] Regulation (EU) 2019/1020 on market surveillance and product conformity.

[63] EUCPN structures the administrative support necessary to integrate and coordinate resources, as well as to facilitate cooperation and information exchange between market surveillance authorities in the Union. It is made up of these authorities of each MS, national experts and the Commission itself. Álvarez García, V., *Las normas técnicas armonizadas (una peculiar fuente del Derecho europeo)*, Iustel, Madrid, 2020, p. 53.

tive implementation of the Regulation, thereby discouraging infringements[64]. Hence, in the framework of the network's activity, the SSMS provides harmonised market surveillance criteria applicable to AI systems, completes the framework for the conduct of coordinated investigations and provides the technical and administrative support structure by acting as a contact and coordination node for the Administrative Cooperation Group (ADCO).[65]

It is therefore an informal group composed of market surveillance authorities and the Commission[66]. Its chairman is appointed by and from among its members, holding regular meetings under the administrative and financial support of the Commission[67]. Its ultimate mandate is to promote the efficient supervision of AI systems subject to the AIA in accordance with the principles of proactive oversight, proportionality, and cooperation[68]. The specific tasks do not differ from those recognised for these groups by the market surveillance rules (Art. 32 MRS), and aim at enhancing the effectiveness of proactive[69] and reactive market surveillance activity. In relation to the latter, in case of incidents, accidents or complaints, which is the scope of reactive market surveillance activity, the AI ADCO Subgroup will make use of the European Information and Communication System for Market Surveillance (ICSMS), a support platform for the exchange of information and coordination of activities between the non-food ADCOs, which will act as a node for reporting incidents in the use of suspected non-compliant or risky AI products and systems.

[64] Cons. 55 of Regulation (EU) 2019/1020 on market surveillance and product conformity.

[65] The integration of the SSMS into EUCPN is implemented through its configuration as an AdCo Group. Therefore, the final design of this sub-group and its functions are set out in the Market Surveillance and Product Conformity Regulation, and not in the AIA which is configured as the *lex specialis* of this regulatory framework.

[66] Art. 11.8 of Regulation 2019/1020.

[67] This funding is provided for in Art. 32.e) of Regulation (EU) 765/2008 of 9 July 2008 setting out the requirements for accreditation and market surveillance relating to the marketing of products.

[68] In line with Art. 30 of Regulation 2019/1020.

[69] With regard to proactive oversight, the AI ADCO Sub-Group will be responsible for promoting uniform application of the AIA and Regulation 2019/1020, fostering communication and trust between MS market surveillance authorities, coordinating joint projects and developing common methodologies to ensure effective supervision, especially in case of cross-border activities. In addition, this mechanism allows exchanging information on best practices and aligning them with the general ones already collected in the field of market surveillance, addressing issues of common interest to propose unified approaches and facilitating sector-specific assessments, including risk analysis and scientific developments. Ultimately, it allows for optimising and streamlining upstream monitoring activities.

For its part, the Standing Sub-Group of Notifying Authorities (SSNA) is charged with cooperation on issues related to notified bodies, without specifying its composition and functions in greater detail. This brief mandate, however, needs to be complemented with European legislation on products, and more specifically with the provisions of the "Blue Guide"[70], which also includes the mandate to establish cooperation instruments.

VIII. Other advisory, support and collaboration bodies

In addition to the governance entities already discussed, the Act introduces additional ones: the Advisory Forum, the Scientific Panel of Independent Experts (the 'scientific panel'), and the European Centre for Algorithmic Transparency (ECAT), together with the AI test support structures. They are designed or linked to integrate diverse perspectives, as well as technical expertise and knowledge in order to promote a harmonious and effective implementation of AI in the single market.

1. Advisory Forum

To respond in a balanced way to the challenges of AI, the expertise of regulators armed with a powerful arsenal of powers and competences is not enough. It is necessary to include and value the perspective of the other actors in the European ecosystem. In this sense, the incorporation of the latter through the Advisory Forum is a spur for the effective implementation of AIA, and to accompany the rapid technological evolution by integrating technical expertise with the different economic and social sensitivities. In this way, the coexistence of the risk-based approach with the principle of innovation can be assessed.

The Regulation's approach to this issue, although more comprehensive, does not introduce new features with respect to the Community product safety framework, which already required the effective representation and participation of stakeholders[71]. The Regulation adopts and extends this collaborative approach, as reflected in its Recital 150 and Art. 67, with the aim of promoting the legitimacy and acceptance of the Community framework.

The Forum should provide additional technical expertise, and is recognised as having the power to prepare opinions and recommendations to

[70] Commission Communication "Blue Guide" on the implementation of European product legislation, (2016/C 272/01) of 26 July 2016.

[71] See, thus, Art. 5 Regulation (EU) 1025/2012.

the CAI and the Commission. Its composition should represent the effort to incorporate the broad spectrum of interests and perspectives of the *stakeholders* of the Union's AI eco-system, incorporating in a balanced way academia, industry, start-ups, SMEs and civil society. This representation seeks to promote a constructive dialogue, addressing the implications for different sectors and constituencies, as well as taking into account commercial, economic and societal interests.

The Commission should appoint the members of the Forum on the basis of their proven expertise in Artificial Intelligence, ensuring diversity. Its term of office, initially two years, may be extended to a maximum of four years. It should have the autonomy to adopt its own rules of procedure and elect co-chairs from among its members. It will meet at least every six months and may invite experts and other actors from the AI ecosystem, although the European Union Agency for Fundamental Rights (FRA), ENISA, and the European standardisation bodies (CEN, CENELEC and ETSI) will have the status of permanent members.

2. Scientific panel of independent experts

The Interim Agreement on Interinstitutional Negotiations introduced the Scientific panel of independent experts (SPIE) to support the application and implementation of the Regulation, especially with regard to general purpose models (Art. 68). It is to be set up and made operational by the Commission by means of an implementing act, following the examination procedure set out in Article 5 of Regulation (EU) 182/2011.[72]

The number of experts shall be determined by the Commission after consultation of the CAI, ensuring equitable geographical and gender representation. They will be selected on the basis, as a minimum, of their proven scientific or technical expertise in AI; their independence from general purpose system or model providers; and their ability to act in a diligent and objective manner. They are subject to obligations of impartiality, objectivity and confidentiality, operating without instructions from third parties in order to preserve their independence. To promote transparency, the AIO will implement specific procedures to prevent conflicts of interest, including the mandatory publication of publicly accessible declarations of interest.

Although the SPIE's range of responsibilities is broader, its role is con-

[72] Regulation (EU) 182/2011 of the European Parliament and of the Council of 16 February 2011 laying down the rules and general principles concerning mechanisms for control by Member States of the Commission's exercise of implementing powers.

centrated on providing technical support and advice on general purpose models (Recital 151 and Article 68.3). This role differs from that attributed to the Advisory Forum, which not only incorporates state-of-the-art technical knowledge, but also encompasses a broader perspective by paying attention to various sectoral sensitivities. In this way, the functions of both bodies complement each other and enrich the governance framework with a solid, comprehensive and multidisciplinary vision.

The functions of the SPIE are instrumental in ensuring the effectiveness of the AIO by providing an indispensable scientific and technical reference framework. Its remit includes issuing qualified warnings to AIO on general purpose models that present systemic risks at EU level. It also develops methodologies to assess both the capabilities and the risks associated with general-purpose AI systems, strengthening the Office's ability to respond to situations requiring safeguard measures.

While the advisory functions are addressed in particular to the AIO, they extend to market surveillance authorities and it is the Commission's task to ensure access to the *pool of experts*, although this may entail the payment of fees and charges (Art. 69).

3. European Centre for Algorithmic Transparency and Artificial Intelligence Testing Support Structures

The European Centre for Algorithmic Transparency (ECAT), established in the framework of the Digital Services Regulation[73], is called upon to actively collaborate with the AIO[74]. Without wishing to exhaust its regulation and functions, it is configured as a support structure of DG Connect integrated in the Commission's Joint Research Centre. Its main task is to provide scientific and technical advice in research on algorithmic systems implemented by online platforms, in order to ensure that they comply with Community law.[75]

ECAT is generally credited with inspecting and developing technical tests of algorithmic systems to understand their performance, identifying and quantifying systemic risks of online platforms and large search engines (VLOSEs), and ultimately developing methodologies to assess the fairness

[73] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a digital single market for services.

[74] This is stated in Art. 5.2.a of the Decision establishing the Office for Artificial Intelligence.

[75] The doctrine has dealt with them. On this subject, see, among others, Ilichman, D., "European Approach to Algorithmic Transparency", *Charles University in Prague Faculty of Law Research Paper*, no. 2023/II/1, 2023, p. 11 et seq.

of algorithmic models. Given this expertise, as well as the fact that some algorithmic models subject to the DSA may apply AI models, collaboration between ECAT and AIO is indispensable. The specific formulas for implementing this cooperation are yet to be defined, *but* the responsibility for developing such mechanisms is attributed to the AIO (Art. 3 of the decision).

Article 84 mandates the Commission to designate one or more Union AI test facilities. Such facilities, which are regulated in Regulation 2019/1020 (Art. 21), support the work of Market Surveillance Authorities, the EUPCN, the Commission and other public entities.

Once designated by the Commission, they are attributed the same general functions as those set out in art. 21.6 of the MRS, but in the field of AI. This is, in our case, to perform tests and evaluations of systems and products or with AI components upon request of the AVMs, the AIO or the Commission; to provide independent technical or scientific advice in the framework of the cooperation that may be developed through EUPCN and the ADCOs; and to develop new techniques and methods of analysis to assess AI risks and models. In addition to this, either the CAI, the Commission or the AVMs may request independent technical or scientific advice from them, if they deem it relevant.

IX. Safeguard, market surveillance and control procedures for Artificial Intelligence systems in the Union

As has already been discussed at various points, the AIA is part of the Union's harmonisation legislation and therefore has to comply with the institutional structure, techniques and procedures provided for by Regulation 765/2008, the MRS and Decision 768/2008.

Although it is well known, it is worth remembering that the anchorage of this power of harmonisation and approximation of EU legislation is contained in Art. 114 TFEU[76], which is also one of the bases of the AIA[77]. This provision allows harmonisation legislation to include and authorise the use of a safeguard clause by the MS (Art. 114.10), on the basis of which they developed specific safeguard procedures, which are now generally regulated by

[76] This has been dealt with in the doctrine. Among others, Barnard, C., *The Substantive Law of the EU. The Four Freedoms,* seventh edition, Oxford University Press, New York, 2022, pp. 557 ff.

[77] The other, as already discussed in another chapter of this paper, is Art. 16 TFEU, which serves as a basis for regulating certain uses of AI involving the processing of personal data.

Chapter III of Regulation 765/2008 and Decision 768/2008[78]. The AIA acts as a *lex specialis* and therefore completes and specifies this general framework in Articles 79-83, which we will examine below.

However, although this is addressed in other chapters of this book, it is worth noting that these safeguard procedures and market surveillance measures are triggered once AI systems deployed or operating in the market but presenting a serious risk that may adversely affect the health, safety or fundamental rights of persons are identified, detected or notified[79]. General Community legislation regulates the cases of serious risks that have a national impact, differentiating them from those that may reach other MS. It also covers cases of compliant products that nonetheless pose a risk to health and safety, and formal non-compliance. All of these are covered by the AIA, which, however, introduces a specific procedure for systems classified by providers as not high risk in application of Annex III.

1. Procedure for Artificial Intelligence systems posing a risk at national level

The market surveillance rules regulate a first procedure in case of non-compliance or incidents limited to the territory of a Member State. The procedure, set out in Art. 79, is specific to AI systems and follows the guidelines set out in Regulation (EC) 765/2008, Art. 19 of the MRS, as well as Art. R31 of Annex I of Decision 768/2008/EC. In this sense, if a market surveillance authority has indications of the existence of a risk, it will initiate the procedure with a view to assessing whether the risks are serious and whether the requirements and obligations of the AIA are not being complied with, especially in the case of the systems covered by Art. 5 that affect vulnerable persons. To this end, they have the aforementioned powers, which are reinforced by Art. 14.3 MRS, and, in particular, the aforementioned powers of access to information, with providers and other operators being obliged to cooperate. These operators, as well as others in the life cycle of the AI system concerned, will be notified of the procedure. In addition, in case of risks to fundamental rights, the national authorities competent to supervise

[78] On these procedures, in general terms, see Álvarez García, V., *Derecho de la regulación económica. VIII. Industria*, Iustel, Madrid, 2010, pp. 474-475.

[79] The definition of an AI system presenting risk is to be understood not only in terms of Art. 79.1 AIA, but also in terms of Art. 3.19 of the MRS. Therefore, for the purpose of defining risk, it has to be considered whether the risk is reasonable and acceptable in view of the intended purpose, normal uses or foreseeable uses.

and ensure the obligations in this area will be informed and will be required to cooperate with them.

As the ultimate rationale of the procedure is to ensure rapid intervention to prevent risks from materialising or to contain their spread, once the risk is established, the relevant operator will be required either to take corrective action to bring it into compliance, or to withdraw[80] or recall it[81] from the market, within a reasonable and proportionate period to be set by the supervisory authority taking into account the nature of the risk, but not exceeding 15 working days or the period provided for in the relevant harmonisation legislation.

If the relevant operator fails to take the measures, the regulation empowers the market surveillance authority to take interim measures in accordance with national law. These measures may be all appropriate and consistent measures to withdraw or recall the system or, as the case may be, to prohibit, or restrict placing on the national market.

Within the framework of the procedure, the rights and guarantees set out in Art. 18 of Regulation 2019/1020 must be respected, so that both the assessment of the serious risk and the measures, as well as the deadline for adopting them by the operator, must be adequately justified, also communicating the corresponding means of appeal in accordance with domestic law. Furthermore, the right to be heard must be guaranteed, either before the adoption of the decision and within a period of no more than 10 days, or afterwards, in the event that the delay in its adoption means that the risk has materialised, and therefore immediate intervention is advisable.

If the non-compliance goes beyond the territory of one Member State, the market surveillance authority, after evaluating and assessing the risk, must inform the Commission and the other Community partners of the outcome of the evaluation and of the corrective measures to be taken by the operator concerned. These measures, moreover, must be adequate not only to respond internally, but also to cover all systems placed on the market in the Union. Again, in the event that the operator does not adopt them, it is empowered to adopt interim measures with the aforementioned scope, again informing the Commission and other MS.

Given the extent of the effects of non-compliance, a more granular exchange of information should be ensured in these cases. Hence, the proposal is responsible for establishing the minimum points to be communicated (Art.

[80] This means any measure aimed at preventing the distribution, display or offer of an AI system.

[81] That is, the adoption of measures to recover a system that has already been made available to users.

79.6). These are: the necessary data to identify the non-compliant system; the traceability and origin; the nature of the non-compliance and the risk, as well as the nature and duration of the national measures taken; and, the arguments put forward by the relevant operator. In addition, the market surveillance authority must specify whether the non-conformity is due to non-compliance with the prohibition of prohibited practices in Art. 5, non-compliance with horizontal obligations for high-risk systems, insufficient harmonised standards or common specifications, and/or non-compliance with transparency obligations of providers and users of general purpose AI systems generating synthetic audio, image, video or text content.

It is worth noting that the initiation of this procedure empowers the market surveillance authorities of other MS to take the corrective measures they deem appropriate for their territory, which they must also communicate to the other MS and to the Commission itself. In line with this, all MS are also obliged to communicate any additional information about the non-compliance of the AI system concerned.

Once the information has been received, the Commission and the market surveillance authorities of the MS generally have three months to present objections to the notified measures in general cases, and thirty days in cases referring to the prohibited practices of Art. 5. In case of objection, the Community procedure referred to below is initiated.

2. Union safeguard procedure

Article 81 of the AI Act regulates the Community safeguard procedure which, in line with the above, maintains the broad outline of the general rules on this matter.

Thus, this stage of the procedure is initiated if, within the time limits set from the notification of the measures taken under a national safeguard procedure, a market surveillance authority raises objections to the measures taken, or if the Commission itself considers that these measures may be contrary to EU law. In such cases, a consultation process is opened under the leadership of the Commission, with the participation of the MS surveillance authorities and the operators concerned, in order to assess the appropriateness and adequacy of the measures taken by the Member State initiating the procedure at national level.

At the end of this phase, the Commission shall decide and notify the justification of the measure or measures in question, within a period not exceeding six months from the notification referred to in Article 79.5, or sixty days in the case of prohibited practices. If it considers the measure to be justified, all MS must ensure its effectiveness by adopting the appropriate restrictive measures, including the withdrawal of the system, with notification to the Commission; otherwise, the market surveillance authority of the Member State that initiated the procedure must proceed to withdraw it, and must also notify the Commission.

In addition to the above cases, it is possible that the Community administration may consider the national measure to be unjustified, not for reasons of non-conformity, but due to deficiencies or inadequate development of harmonised standards or, if applicable, of common specifications. In this case, the procedure provided for in Art. 11 of Regulation 1025/2012 on European standardisation would be triggered. To this end, the European standardisation bodies must be notified and the Standing Committee of MS representatives must be informed and consulted. The committee will proceed to decide, after appropriate consultations with the standardisation bodies.

3. Procedure in respect of compliant Artificial Intelligence systems presenting a risk

In line with the procedures explained above, the general product safety legislation regulates the procedure for compliant systems presenting risks. Specifically by the aforementioned Decision 768/2008/EC. Article 82 of the regulation does not deviate substantially from the provisions of this decision, and to a large extent replicates the procedural *iter* commented on with regard to the Community safeguard procedure. This is why we will only mention the differential aspects.

The procedure also applies when a market surveillance authority identifies a risk to the health or safety of persons or a breach of obligations under Community or national law to protect fundamental rights or other overriding public interest objectives, and informs the Commission and the other Member States[82]. However, the assessment carried out by the national authority shows that the system complies with the harmonised technical standards and the provisions of the Regulation.

In this case, it must consult with the competent authority on fundamental rights, in accordance with Art. 77.1. It must also request the operator to adopt the appropriate corrective withdrawal or recovery measures, set by the national authority, which must apply to all systems marketed in the EU. Sub-

[82] The notification must specifically provide the data required to identify the affected AI system, as well as to determine its origin and supply chain. It should also state the nature of the risk presented and describe the nature and duration of the measures implemented.

sequently, the Commission and the MS will be notified, the aforementioned consultation process will be carried out, and it will be the Commission that will decide on the justification and suitability of the measures adopted, notifying the MS and the economic operators affected.

4. Procedure in case of formal non-compliance

Finally, Article 83 of the AI Act contains a procedure aimed at responding to non-compliance with certain formal obligations: conformity marking that does not comply with the specifications set out in Article 48; non-existence of conformity marking; failure to draw up or incorrect drawing up of the EU declaration of conformity; failure to designate an authorised representative, where applicable; failure to provide technical documentation; or failure to register in the EU database, in accordance with Article 71.

In all such cases, the market surveillance authority shall require the relevant provider to remedy the non-compliance and, in the event of refusal or persistence, shall take appropriate measures to restrict or prohibit the placing on the market or, where appropriate, its recall or withdrawal from the market without undue delay.

5. Procedure for dealing with AI systems classified by the provider as non-high-risk in application of Annex III

Article 80 of the AIA establishes a novel procedure for assessing the classification of AI systems that providers consider not to be of high risk. As can be inferred, the objective is to ensure that the classification made by providers is appropriately adjusted to their level of risk, taking into account the criteria set out in Art. 6.3 and the corresponding guidelines developed by the Commission.

In line with the above-mentioned cases, it is up to the market surveillance authority to assess the system. If it finds that it has been wrongly qualified, it will require the provider to comply with the requirements and obligations established for high-risk systems, and to adopt the relevant corrective measures within a reasonable period of time to be set by the authority. Also along the same lines, if the use of the system in question exceeds the territorial scope of the market surveillance authority, it must notify the European Commission and the Member States, reporting both the assessment and the measures taken.

Once notified, the provider must take measures to comply with the requirements and relevant obligations to the high-risk system concerned. It must also ensure compliance with the corrective measures which, where applicable, must cover all systems placed on the market in the Union territory.

In the event of non-compliance with such requirements, sanctioning procedures will be opened which may lead to the imposition of sanctions provided for in Art. 99. The same dissuasive measure is taken in the event that it is determined that the classification of the system as not high risk by the provider was intended to circumvent the requirements established by Art. 8 to 15 of the AI Act.

THE SANCTIONING REGIME IN THE ARTIFICIAL INTELLIGENCE ACT

F. Javier Sempere

Director of Supervision and Data Protection of the General Council of the Judiciary.
PhD candidate at CEU International Doctoral School (CEINDO).

I. Introduction

The AIA devotes only two precepts to regulating its system of penalties, to which must be added the corresponding recitals explaining their content. These are Articles 71 and 72, entitled "Penalties" and "Administrative fines" respectively for Union institutions, bodies, offices and agencies. These are two precepts whose purpose is different, since the former is the one that actually regulates the sanctioning regime, while the latter is devoted to conferring the sanctioning power within the framework of the Community institutions to the European Data Protection Supervisor (EDPS).

Both, but above all, the first, in terms of the regulatory technique used, is closely related to the articles that, in turn, regulate the sanctioning regime of the European Data Protection Regulation (GDPR)[1], also suffering from the same shortcomings, such as the deficient classification of offences, which is why we will refer to it on occasions.

Furthermore, we must take into account the link between Artificial Intelligence systems and the personal data protection, which will mean, as we will explain, the necessary communication between the AI Supervisory Authority and the Data Protection Supervisory Authority when both are different, given that some countries have determined that the Artificial Intelligence authority

[1] OJEU of 4 May 2016.

https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=celex%3A32016R0679

Subsequently, corrections of errors were published on 23 May 2018 and 4 March 2021. In some cases, these are more of a material correction than an error correction, changing certain aspects of the content of the regulation. It should be borne in mind that the first is made two years after the first publication of the regulation, and therefore, although it was not yet applicable, it had been subject to detailed analysis. And the second, once the standard has been applied.

They can be consulted at

https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32016R-0679R(02)&from=EN

https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32016R-0679R(03)&from=EN

will be the personal data protection authority, without the need to create a new entity.

In addition to these two precepts, some others that appear in the rest of the text of the AIA should also be considered, such as article 66 on "Tasks of the Board", article 70 on "Designation of national competent authorities and single points of contact", and article 79 "Procedure at national level for dealing with AI systems presenting a risk". We will also refer to them in more detail because of their relation to the content of the aforementioned article 71.

With regard to the recitals, although only recital 84 is dedicated to the sanctioning part, for completeness it is also possible to refer to recital 79, which refers to the functions of the so-called market surveillance authorities.

Consequently, we note that the sanctioning regime is, at first sight, rather brief with only two precepts, one of them dedicated to the EDPS, and an explanatory recital, to which a second one could be added. Probably, in addition to some shortcomings, as we have already mentioned, the reason for this is that each Member State must, in part, develop this sanctioning regime.

Likewise, and without prejudice to the content of these precepts, as well as the regulation adopted in the future to implement this development, when exercising the sanctioning power by the market surveillance authority, which, in our case, will be carried out by the Spanish Agency for the Supervision of Artificial Intelligence (AESIA), the principles regulated in Law 40/2015, of 1 October, on the Legal Regime of the Public Sector (LRJSP)[2] must be complied with, such as legality (art. 25), non-retroactivity (art. 26), typicality (art. 27), liability (art. 28), proportionality (art. 29), prescription (art. 30), and concurrence of sanctions (art. 31). These principles have their origins in criminal law, but the Constitutional Court in its ruling 18/1981, of 8 July[3], pointed out that their main purpose is to provide procedural guarantees.

II. Article 71 of the Regulation: the need for legislative development and interaction with other legislation and alternatives to fines and specifications for administrations.

We begin with this provision, to which we will devote a more detailed and detailed analysis, as it is clear from its content that it regulates the sanctioning regime applicable in this area.

[2] BOE n.º 236 of 2 November 2015. https://www.boe.es/buscar/act.php?id=BOE-A-2015-10566

[3] Rebollo Puig, M., Izquierdo Carrasco, M., Alarcón Sotomayor, L., y Bueno Armijo, A. Mª., *Derecho Administrativo sancionador*, Editorial Lex Nova, First Edition, Valladolid, 2010.

We start from the initial text of the European Commission's proposal for AIA, and can distinguish several sections: the referral to the Member States to develop the system of penalties, including whether or not they determine the possibility of imposing fines on the public sector; the classification of infringements and the amount of fines; the criteria for determining the amount of the fine to be imposed when an infringement is committed; and the power to impose fines not only by administrative bodies but also by the courts, depending on the applicable legal system in each country. Let us proceed to analyse each of these sections in detail.

With regard to the first, it means that the system of penalties is not exhausted by the provisions of Article 71, but that the Member States "*shall lay down the rules on penalties and other enforcement measures, which may also include warnings and non-monetary measures, applicable to infringements of this Regulation by operators, and shall take all measures necessary to ensure that they are properly and effectively implemented*". In other words, it will be necessary for each of the Member States to develop legislation to supplement the provisions of the aforementioned provision.

Thus, it can be seen that with regard to how to respond to the commission of an infringement by the supervisory authority, only the possibility of imposing fines has been contemplated, with no other type of alternative such as a warning, a caution, or an order to comply, which is contemplated in the GDPR[4]. It should be remembered that the warning, according to the amendment of Organic Law 3/2018, of 3 December, on the protection of personal data and guarantee of digital rights (LOPDGDD), lacks the nature of a sanction[5]; the warning can be applied when it is possible that an infringement has been committed but without the corresponding sanctioning procedure being processed; and the compliance order, although the Spanish Data Protection Agency (AEPD) applies it in sanctioning resolutions, such as those concerning video surveillance, in which it orders, for example, that the camera does not record the public road or that the right to information is complied with by means of the corresponding sign, other Authorities use this option without the need to process a sanctioning procedure.

In this respect, it should be noted that a fine does not necessarily have to

[4] See in this respect Article 58.2 of the GDPR.

[5] See in this regard Law 11/2023, of 8 May, on the transposition of European Union Directives on the accessibility of certain products and services, migration of highly qualified persons, taxation and digitalisation of notarial and registry actions; and amending Law 12/2011, of 27 May, on civil liability for nuclear damage or damage caused by radioactive materials; amending certain articles of the LOPDGDD. BOE no. 110 of 9 May 2023.

be imposed in the event of a possible infringement, but rather, depending on the specific case, the harm caused can be resolved without the need to resort to the binomial of opening a sanctioning procedure to impose a fine. In fact, the AIA proposal itself, although as we have pointed out, only includes a fine in Article 71, from another of its precepts, Article 79, this possible form of action can be deduced.

Thus, according to it, a specific procedure is established in the case of AI systems that present a risk at national level with respect to health, safety or the protection of fundamental rights, so that the supervisory authority, when it becomes aware of it, does not necessarily have to issue the agreement to initiate the sanctioning procedure, but will carry out an evaluation of the system to verify that it meets all the requirements. And if it does not meet them, it will demand immediate corrective measures, or withdraw it from the market, and may also adopt provisional measures to prohibit or restrict the marketing of the system.

With regard to the interaction with other regulations and alternatives to the fine, this Article 79 also breaks down measures that can also be described as punitive, such as the withdrawal, prohibition, or restriction of a certain AI product. It should also be borne in mind that it is not necessarily the fine that causes the greatest harm, but rather the aforementioned. Withdrawal would temporarily leave without the possibility of income due to the AI product not being available on the market; restriction would mean that income would be lower; and the highest sanction would be to prohibit the product in question.

Consequently, this development could include all these measures that can be adopted by the market surveillance authority, similar to what the GDPR provides for when it regulates the powers of the supervisory authorities. To recapitulate, these could be warnings, cautions, compliance measures, as well as the withdrawal, prohibition and restriction of an AI product.

With regard to the specifications for public administrations, this future regulation should include whether public administrations can be fined in the event of committing one or more infringements, given that the draft AIA leaves it up to each of the EU member states to decide in this regard. This means that there will probably be no uniformity, as some countries may contemplate this possibility and others may not, as has happened with the GDPR, which contains a similar provision, and in which the general rule has been that fines can be imposed, except in three countries, namely France, Luxembourg, and Spain. Therefore, it is most likely that our legislator will adopt the same provision as the one currently in force in the LOPDGDD, recently amended, to replace the possibility of warning public administrations with the system of the former Organic Law 15/1719, of 13 December, on the Protection of

Personal Data, consisting of pointing out that one or more infringements have been committed, the possibility of urging compliance with measures and, where appropriate, although in practice it is a "rara avis", proposing the initiation of disciplinary proceedings against the alleged offender.[6]

In this sense, we do not agree that this possibility is not included, since, on the one hand, it represents a comparative disadvantage with respect to the private sector, and on the other hand, as has been demonstrated in the field of data protection, it represents a "relaxation" in compliance in the public sector. It should be added that in our legal system there are other regulations which, on the other hand, do provide for this possibility, such as some anti-trust regulations[7], or more recently in Law 2/2023, of 20 February, regulating the protection of persons who report regulatory infringements and the fight against corruption.[8]

In any case, if the purpose of a Community Regulation is to ensure that all those affected by its content comply with the same rules, with this type of authorisations in favour of the Member States, the desired uniformity is lost, and we wonder why a Town Council in Italy can be fined if it fails to comply with the AIA and a Town Council in any Autonomous Community, in the same situation, will only receive a warning?

III. Prescription and classification of infringements and amount of fines in Article 71

On the other hand, Spanish law can also cover other procedural issues, but above all, it can regulate the statute of limitations for infringements and penalties. It should be remembered that the statute of limitations is a fundamental element in our legal system and provides legal certainty, especially for

[6] See in this respect Article 77 "Regime applicable to certain categories of controllers and processors" of the LOPDGDD.

[7] Ortega Fernando, J. *Comentario a la STS de 18 de julio de 2016*, in the legal blog almacenderecho.org, 2017.

https://almacendederecho.org/cuando-se-puede-sancionar-la-administracion-ambito-la-defensa-la-competencia

[8] BOE no. 44 of 21 February 2023.

https://www.boe.es/buscar/act.php?id=BOE-A-2023-4513

This regulation has not contemplated a specific regime for public administrations, since according to Article 62.1 *"Natural and legal persons who carry out any of the actions described as infringements in Article 63 shall be subject to the sanctioning regime established in this law"*, with the difference that, in terms of amount, Article 65.1 in section a) establishes a maximum of 30,000 € for natural persons and in section b) 1,000,000 € for legal persons.

the alleged offender, in the sense that, once the offence has been committed, the statute of limitations period begins to run, so that, if it has elapsed, the offender cannot be sanctioned. Article 30 of the LRJSP regulates the statute of limitations for both infringements and sanctions, based on the general rule that both prescribe according to the provisions of the laws that establish them, and failing this, in the case of very serious infringements, three years, serious infringements two years, and minor infringements six years, and in the case of sanctions, the period is the same except for those imposed for minor offences, which will be one year.

In the case in question, the AIA proposal obviously does not contemplate a statute of limitations for offences or penalties, but, above all, and as we will explain below, it does not classify offences as minor, serious, or very serious (nor do the penalties), since it uses a system of classification in "broad terms", in the same way as the GDPR, which can also be described as exhaustive.[9]

Therefore, it will have to be Spanish law that covers the non-existence of the statute of limitations, but closely linked to the classification of offences, and in a similar manner to the LOPDGDD[10], which leads us to analyse the second noteworthy section of this Article 71, consisting of the aforementioned classification, as well as the amounts of the fines to be imposed. Specifically, points 3, 4, and 5 of the aforementioned provision, which we shall now analyse.

As for point 3, it determines that any violation of both the prohibition of Artificial Intelligence practices in Article 5 and the requirements of Article 10 can be fined up to 30 million euros or in the case of an undertaking up to 6% of the annual turnover of the preceding financial year, whichever is higher.

Article 5, as can be seen from its title "Prohibited AI practices", contains a whole list of practices that cannot take place due to this prohibition. This list appears in paragraph 1 of this provision, consisting of the introduction to the market, putting into service or use of an AI system "*which uses subliminal techniques that transcend the consciousness of a person to substantially alter that person's behaviour in a way that causes or is likely to cause physical or psychological harm to that person or to another person*", the purpose of which is to "*assess or classify the reliability of natural persons over a period of time on the basis of their social behaviour or known or predicted personal or personality characteristics*" and which may result in detrimental treatment; or the use of real-time remote biometric identification

[9] Hernández Corchete, J.A., *Exhaustividad y estándares del principio de legalidad sancionadora*, in Derecho Digital e Innovación, n.º 2, Editorial Wolters Kluwer, April-June 2019.

[10] See Articles 72, 73 and 74 which regulate the statute of limitations for very serious, serious and minor infringements.

systems at access points in public areas where one of the exceptions provided for does not apply.

However, in some of the other sections of the aforementioned Article 5, it is also possible to infer the existence of conduct which, in the event of non-compliance, could be subject to sanctions. Thus, section 2 indicates that in the case of a real-time biometric identification system for a public access space for the purposes contemplated in the law, it must comply with the necessary safeguards and conditions regarding its use and particularly with respect to temporal, geographical and personal limitations; and section 3 obliges that prior authorisation by the judicial or administrative authority must be obtained before its operation. Consequently, infringements could be committed by violating its temporal, geographical, or personal limits, or by putting it into operation without having obtained the aforementioned authorisation.

Therefore, although the rule refers only to the list of prohibitions, the two sections mentioned above also contain other conducts that could be subject to infringement in case of non-compliance.

Article 10, entitled "Data and data governance", sets out the quality criteria for the data set to be used for training, validation, and testing of high-risk AI systems. These quality criteria are broken down in paragraphs 2 to 5 of this provision and should be differentiated according to the content of each of them.

Thus, paragraph 2 refers to good governance and data practices, including choosing an appropriate design, data collection, or screening for bias; paragraph 3 requires that training, validation and test data be relevant, representative and error-free, with appropriate statistical properties, which may be for each data item or combination of data items; and paragraph 4 mandates that training, validation and test data take into account particular characteristics or elements of the specific geographic, behavioural or functional context; paragraph 4 mandates that the training, validation and test data take into account the particular characteristics or elements of the specific geographical, behavioural or functional context; and finally, paragraph 5 enables providers of these high-risk systems, where strictly necessary to ensure the monitoring, detection and correction of the associated biases, to treat special categories of art. 9.1. of the GDPR as well as art. 10.1 of Regulation 2018/1725[11], provided that adequate safeguards for the fundamental rights and freedoms of

[11] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of individuals with regard to the processing of personal data by Union institutions, bodies, offices and agencies, and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC. OJEU of 21 November 2018.

natural persons are provided, including establishing technical limitations to the reuse and use of recent security and privacy protection measures, such as pseudo-anonymisation or encryption, where anonymisation could significantly affect the intended purpose.

In view of the content of this Article 10, we must conclude that there does not necessarily have to be non-compliance with all the requirements in order to find an infringement, but that the lack of one of them would be punishable. Therefore, it could be, for example, that with regard to paragraph 2, an adequate design and data collection had been carried out, but the examination had been omitted due to possible biases; with regard to paragraph 3, the training, validation, and test data lacked errors but were not relevant; with regard to paragraph 4, the geographical and behavioural context had been taken into account but not the specific context; and with regard to paragraph 5, technical limitations to re-use had been adopted by establishing security measures but not all of them had been necessary. As can be seen from these examples, if some requirements are met but others are not, non-compliance could result in the commission of the respective infringement, leading to a sanction, which would be a fine.

Likewise, with regard to the last paragraph, a dual interpretative task will have to be carried out for its application, since, on the one hand, this authorisation is limited to being "strictly necessary", so it could happen that it was not and such data had been used; and on the other hand, since this authorisation concerns personal data protection, the competent authority will probably be the Data Protection Authority, and not the AI Supervisory Authority. This interaction with data protection regulations, not only in this provision but also in others, should be clearly defined in the text so that it does not lead to future conflicts[12].

In short, from these two precepts multiple non-compliances can be deduced, so it will be necessary to go to each of them to break them down in order to be able to determine which conducts are punishable. However, as we have described, they refer to a list of high-risk activities (Article 5) and to requirements (Article 10), so that, as far as possible, this action can be carried out without inconvenience, although it would have been more protective if the regulation had contained a list of conduct liable to be fined or, where appropriate, sanctioned with another figure.

[12] CEPD-SEPD: *Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act).* 2021. https://edpb.europa.eu/system/files/2021-10/edpb-edps_joint_opinion_ai_regulation_es.pdf

However, such a list should exist, especially if we take into consideration the total lack of legal certainty caused by the following paragraph, the fourth, of Article 71, which stipulates that any "*breach by the AI system of any of the requirements or obligations laid down in this Regulation other than Articles 5 and 10 shall be subject to administrative fines of up to 20 million euros, or if the offender is an undertaking, up to 4% of the total annual worldwide turnover in the preceding financial year, whichever is higher*". In other words, this means having to review each of the articles of the rule and try to elucidate where there may be a breach and therefore liable to impose a fine, and where not because the content of the corresponding article and its paragraphs are merely declaratory.

Thus, for example, in Article 11 on "Technical documentation", which contains three paragraphs, obligations arise from the first two, but not from the third. Thus, according to the first, a technical documentation must be drawn up prior to the placing on the market of the AI system and updated; according to the second, when placing on the market or putting into service an AI system listed in Annex II, Section A, a single technical documentation must be prepared containing all the information stipulated in Annex IV, as well as the information required by those legislative acts; and according to the third paragraph, no obligation is contained, since its content implies a reservation in favour of the Commission to adopt delegated acts.

This work, as we have explained, must be carried out on all the precepts of the regulation, although there may be a greater impact in Title III on High-Risk Systems, which includes Chapter 3 "Obligations of providers and deployers of high-risk AI systems and other parties".

As for paragraph 5 of this Article 71, the third block of infringements concerning the "supply of incorrect, incomplete or misleading information to notified bodies or national competent authorities in reply to a request" may be fined up to EUR 10 million or, if the offender is an undertaking, 2% of its total worldwide annual turnover for the preceding financial year, whichever is higher.

Once again, a "titanic" task must be carried out by reviewing those articles of the European standard where there is an obligation to provide information to the aforementioned notified bodies and national competent authorities, which, if not carried out correctly, would be liable to a fine. Thus, for example, the aforementioned Article 11 on technical documentation states that the national competent authorities and notified bodies shall be provided with all the information necessary to be able to assess whether the AI system meets the corresponding requirements; or Article 22 entitled "Information obligation", which, as its name suggests, places an obligation on the provider, so that if the provider becomes aware that the high-risk AI system presents a risk within the meaning of Article 79 ("Procedure at national level for dealing with AI systems presenting a risk"), he must immediately inform the national authorities of the risk and the corrective measures taken.

Likewise, and in addition to the information obligations set out in the regulation, it is also necessary to consider the information requests within the framework of an investigation, which usually take place before the initiation agreement is issued, in the preliminary proceedings phase, and whose purpose is to ascertain whether there is sufficient evidence to consider the possible commission of one or more infringements, and, consequently, to issue the initiation agreement. Failure to reply to these requirements, or a reply that does not cover everything that is required, could also be punishable.

On the other hand, with regard to the amounts of the three sections of Article 71 described above, if we take into account that they range from 30 million or 6% (section 3 on breaches of Articles 5 and 11); 20 million or 4% (section 4 for the remaining breaches); and 2% (section 5 for those arising from providing inaccurate, incomplete or misleading information), for the purposes of the future Spanish law regulating the statute of limitations, based on these amounts, the classification between very serious, serious and minor infringements would have been made. Moreover, in the same way as the GDPR did in its day, this lack of classification of infringements could be "fixed" by means of the statute of limitations, describing the conducts for this purpose, but which may help to know which are subject to sanction.

IV. Criteria for determining the amount of the fine under Article 71

Continuing with the amounts, in order to determine the fine to be imposed, certain factors or criteria must be used in this respect, which can act as aggravating or mitigating factors and which appear in paragraph 7 of this Article 71, and which involve applying the principle of proportionality, which will be violated if the circumstances that motivate the imposition of a fine are not specified[13].

Let us bear in mind that our legal system, in the LRJSP, also includes it in Article 29, these being the degree of culpability or the existence of intentionality; the continuity or persistence of the infringing conduct; the nature of the damage caused; and recidivism, for committing more than one infringement

[13] Hernández Jiménez, H. M. *Aplicación práctica de los principios de la potestad sancionadora de la Administración en la nueva Ley 40/2015*, in Actualidad Administrativa, n.º 2, 2017.

of the same nature within a period of one year when this has been declared by a final administrative decision.

The draft of the AIA, on the other hand, only considers as criteria the nature, gravity and duration of the infringement and its consequence; whether other market surveillance authorities have already imposed administrative fines on the same operator for the same infringement; and the size and market share of the operator committing the infringement. We proceed to explain each of these.

As regards the nature, gravity and duration of the infringement and its consequences, this involves assessing what type of infringement is involved in terms of the three existing categories in relation to paragraphs 3, 4 and 5, since the amounts of the fines are different and, in turn, linked to the nature of each of them; Thus, if the infringement is of paragraph 3, the amount will have to be higher than if it were of paragraph 5; it also means taking into account the damage caused to those possibly affected; also the number of people affected and the duration, so that the more people there are and the longer it lasts, the higher the amount will have to be.

With regard to whether other market surveillance authorities have already imposed administrative fines on the same operator for the same infringement, we must consider that the fine has been imposed by another authority in another country, which necessarily implies that there is communication between the different authorities. In this respect, we must make two clarifications: on the one hand, this criterion does not imply that they will not be sanctioned because they have already been sanctioned, since if this were the case, it would have been expressly contemplated; and on the other hand, it could be interpreted as an aggravating element, and to a certain extent related to recidivism, since if that operator, which operates in several EU countries, has already been sanctioned in one of them for an AI product, it should have corrected the facts that led to that sanction, so as not to cause further damage.

As regards the size and market share of the operator, it is clear that it will be assessed whether it is a small or medium-sized enterprise or a larger one, and that to a certain extent, it appears in the amounts of the fines when referring to the percentages of the turnover of the companies, in line with paragraph 1 of this Article 71 which refers to considering in the penalties to be imposed "*the interests of SMEs, including start-ups, and their economic viability*". Obviously, the larger the size and market share, the higher the amount.

It is also worth mentioning that although these are the only criteria contemplated in the AI draft, through this Spanish rule establishing other sanctioning measures, as well as the statute of limitations, it could also include others to be assessed, such as culpability (if there is intent or negligence), the possible benefit obtained, recidivism, the measures adopted to mitigate the damage caused, cooperation with the supervisory authority, or if in the case of causing damage or harm to persons, if they are minors.

This is without prejudice to the role attributed to the European AI Committee in Article 58 to adopt guidance documents including guidelines for setting administrative fines.

Finally, as regards the last paragraph of Article 71.9, given that some countries do not impose fines by their administrative bodies, it leaves open the possibility of fines being imposed by courts and tribunals. This provision, like others, is also included in the GDPR.

Subsequently, in the text adopted at the European Council of 6 December 2022, the common position on the AIA draft, proposed to introduce numerous amendments, which can be grouped into two different groups, such as greater consideration for small and medium-sized enterprises, and to provide the penalty system with greater legal certainty, including in this case some provisions that we had previously pointed out as necessary.

With regard to the first group, paragraphs 3, 4 and 5 of Article 71 provide that, in the case of small and medium-sized enterprises, the maximum limit of the amount to be imposed in the event of an infringement would be lower. Thus, in paragraph 3, against 6% of the total annual volume of the lower financial year, if SMEs, especially start-ups, are concerned, the limit is set at 3%; in paragraph 4, from 4% to 2%; and in paragraph 5, from 2% to 1%. However, for practical purposes, this limit on downgrades might not have been necessary and the previous limits could have continued to be applied with the graduation criteria, that referring to the size of the company and market share, as well as the provisions of paragraph 1, which contains a mandate to also consider SMEs.

It should be remembered that in its initial wording it stated that any infringement not provided for in Article 3, which in turn established the infringements of Articles 5 and 10, would be fined up to 20 million euros or 4% of turnover, which meant revising the rule from top to bottom in search of possible breaches of any of its precepts.

However, with the new wording, in order to provide greater legal certainty, there is a whole list of provisions susceptible to possible breaches, such as breaches of the obligations of providers under Articles 4b (“requirements for general-purpose AI systems and obligations of providers of such systems”) and 4c (“exceptions to Article 4b”); the obligations of providers under Article 16 (“obligations of providers of high-risk AI systems”); the obligations of other persons under Article 23a (“obligations for other persons to be subject to the obligations of a provider”); the obligations of professional representatives under Article 25 ("professional representatives"); the obligations of importers under Article 23 ("obligations of importers"); the obligations of distributors under Article 24 ("obligations of distributors"); the obligations of users under Article 29(1) to (6a) ("obligations of deployers of high-risk AI systems"); requirements and obligations for notified bodies under Article 31 ("requirements relating to notified bodies"), Article 33(1), (3) and (4) ("subsidiaries of notified bodies and subcontracting"), and Article 34a ("operational obligations of notified bodies"); and transparency obligations for providers and users under Article 50 ("transparency obligations for providers and deployers of certain AI systems"). In this way, it is at least made clear which articles of the text are liable to be sanctioned in the event that their content is violated. However, despite this clarification, it will be necessary to analyse each precept to determine when there is an obligation that can be breached and when there is not, as nothing is imposed and the wording is merely enunciative. It should also be pointed out that the possible infringements of Article 10 which were covered by Article 71.3 have also disappeared, leaving only those relating to Article 5.

Likewise, although there is still no section clearly stating who the possible parties responsible for an infringement may be, it is clear from the list of articles referred to that they would be providers, representatives of providers established outside the European Union, deployers and notified bodies.

Continuing with this second group, greater legal certainty is also provided by the introduction in paragraph 6 of new criteria for quantifying the amount of the fines to be imposed, which, as we explained above, in the draft AI were rather brief. In fact, some of the criteria that we emphasised should also be included, such as intentionality or negligence in the infringement, and any measures adopted by the operator to remedy the infringement and mitigate the possible adverse effects of the infringement.

A third criterion is also included, which provides for the possibility of assessing whether that operator has been fined by other authorities for infringements of national or EU law, where such infringements arise from the same activity or omission that constitutes a relevant infringement of the AIA. Admittedly, this criterion is cumbersome to draft, since it implies that a given act constitutes an infringement of the rule, which means analysing the content of the text of the AIA, looking for where there may be infringements, which, in turn, could be infringements of another matter, regulated by national legislation of the Member States or of the Union. From our point of view, this confluence could occur in the field of personal data protection, when any of the provisions of the GDPR expressly mentioning circumstances or obligations regarding the processing of personal data are infringed.

On the other hand, and in addition to the changes already mentioned, two further changes have been added, consisting of the inclusion in the first paragraph of Article 71 of the provision for the use of the AI system in the context of a non-professional personal activity, and a new paragraph 10 guaranteeing that the market surveillance authority acts in accordance with the procedural guarantees of Union and Member State law, including effective judicial protection.

Regarding such non-professional personal activity, Article 2.8 in fact exempts the application of the AIA when this situation arises, except for the provisions of Article 52 which contains transparency obligations for users of both a biometric categorisation system and an emotion recognition system as well as a system that generates or manipulates images, sound or video that resembles persons, objects, places or other existing entities or events and which may mislead the public into believing that they are genuine or truthful.

This non-application, apart from the aforementioned exception, is very reminiscent of the so-called "domestic exception" of the GDPR[14], which implies its non-application when activities carried out by a natural person, such as the address book of a mobile phone or email, both of which are private and non-professional, can be qualified as such. The personal data protection supervisory authorities interpret this exception in a very restrictive way, so that, for example, a publication of a photograph or video of sexual content of a third party on a social network would mean that the person who has published it could be sanctioned for infringing the GDPR, and more specifically, for having published it on that channel in such a way, allowing indiscriminate access, and there being a lack of legitimisation to do so.

In any case, the provision of Article 71.1 must be understood in the circumstances of Article 52, otherwise it would be meaningless, since for the rest of the cases the AIA does not apply.

With regard to the procedural guarantees for action by the market super-

[14] See in this respect:

COURT OF JUSTICE OF THE EUROPEAN UNION, Case C-101/01. Bodil Lindqvist and Göta hovrätt. 6 November. ECLI:EU:2003:596, 2003.

https://curia.europa.eu/juris/document/document.jsf?docid=48382&doclang=ES

COURT OF JUSTICE OF THE EUROPEAN UNION, Case C-212/13. František Rynes and Urad pro ochranu osobnich udaju. 11 December. ECLI: EU: C:2014:2428, 2014.

https://curia.europa.eu/juris/document/document.jsf?docid=160561&doclang=ES

COURT OF JUSTICE OF THE EUROPEAN UNION, Case C-25/17. Tietosuojavaltuutettu with the intervention of Jehovan todistajat - uskonnollinen yhdyskunta. 10 July. ECLI:EU:C:2018:551, 2018.https://curia.europa.eu/juris/document/document.jsf?docid=203822&doclang=ES

visory authority, these extend to the administrative procedure to be followed in the processing of the corresponding sanctioning proceedings, in which the allegedly responsible party may make allegations and present evidence, as well as the possibility that once a sanction has been handed down in administrative proceedings, the sanctioned party may appeal to the courts.

V. Proposals and changes by Parliament and in the final versions

1. Proposed changes in Parliament

Following this text adopted at the European Council of 6 December 2022, the common position on the draft AIA, on 14 June 2023, also proposes to introduce numerous amendments within the European Parliament, considerably affecting the content of Article 71, the most relevant being the following.

Firstly, the power granted to EU countries to determine the sanctions regime would be limited to infringements committed by any operator, taking into account that this figure includes, according to its definition, *"a provider, product manufacturer, deployer, authorised representative, importer or distributor"*.

Secondly, it increases the content of the fine to be imposed under Article 71.3, on breaches of the prohibition of Artificial Intelligence practices in Article 5, to EUR 40 million or 7% of the total annual worldwide turnover of the previous financial year, whichever is higher; a new paragraph 3a provides that breaches, in addition to Article 10, of Article 13, may be fined up to EUR 20 million or 4% of turnover; breaches of the remaining provisions, other than Articles 5, 10 and 13, shall be fined up to 10 million euros or 2 euros of turnover; and in paragraph 5 of the aforementioned provision, with regard to breaches concerning inaccurate, incomplete or misleading information to notified bodies and national authorities, the fine is reduced to EUR 5 million euros or 1% of its total annual turnover.

Thirdly, there are other instruments derived from the exercise of the sanctioning power, to which we alluded, which could be regulated by the internal regulations of each country, such as orders or warnings, which could be used instead of fines.

Fourthly, more criteria are introduced to assess the amount of the fine, most of which are already provided for in the GDPR, such as "the actions taken by the operator to mitigate the harm or damage suffered by the persons concerned"; "intent or negligence"; "the degree of cooperation with the competent national authorities in order to remedy the infringement and mitigate its possible adverse effects"; "the degree of responsibility of the operator, taking into account the technical and organisational measures it implements"; "the way in which the competent national authorities became aware of the infringement, in particular whether the operator notified the infringement and, if so, the degree of cooperation with the competent national authorities"; "the degree of responsibility of the operator, taking into account the technical and organisational measures it implements"; "the manner in which the competent national authorities became aware of the infringement, in particular whether and to what extent the operator notified the infringement"; "adherence to approved codes of conduct or certification mechanisms"; "any other relevant previous infringements by the operator"; and "any other aggravating or mitigating factors applicable to the circumstances of each case".

For the content of each of them, in the field of personal data protection, but which could give an idea of their application to breaches by AI providers, see the documents approved by the EDPB on this subject.[15]

Fifthly, it establishes the obligation to annually report to the AI Office on the fines that have been imposed, which is constituted as an independent body of the European Union, being able to adopt guidelines, jointly with the Commission, on the rule, and which must be taken into account for sanctioning purposes.

And sixthly and lastly, a paragraph 8 bis is added to Article 71 prohibiting that penalties, litigation costs, and compensation claims may not be the subject of contractual clauses or any other form of burden sharing between providers and distributors, importers, deployers, or any third party. On this prohibition, and its application to the field of personal data protection, our Supreme Court has ruled that the breaches of a controller cannot be attributed to its processor so it is the latter who pays the fines or compensation.[16]

2. Developments in the final texts

In order to finalise the text, during the Spanish Presidency of the Council, and for three days at the beginning of December 2023, trialogues were held to

[15] See in this respect: EUROPEAN DATA PROTECTION BOARD, *Guidelines 04/2022 on the calculation of administrative fines under the GDPR, Version* 2.0. 24 May 2023. https://edpb.europa.eu/system/files/2023-06/edpb_guidelines_042022_calculationofadministrativefines_en.pdf

ARTICLE 29 DATA PROTECTION GROUP, *Guidelines on the application and setting of administrative fines for the purposes of Regulation 2016/679,* 3 October 2017.

https://ec.europa.eu/newsroom/article29/items/611237/en

[16] See in this respect SPANISH SUPREME COURT, SALA DE LO CIVIL, STS 1543/2023 - ECLI:ES:TS:2023:1543, 19 April 2023.

finalise the text, which, after some adjustments, in its version, although not yet published in the OJEU, was published at the beginning of February 2024.[17]

In this final version of Art. 71, compared with the other versions previously presented, the changes in the amounts of the fines to be imposed for the breaches that may occur, as well as the criteria for grading the infringements, with some of those proposed disappearing.

Thus, on the amounts to be imposed in the event of non-compliance, we find the following scale:

- The highest fine, up to €35 million or 7% of its total worldwide annual turnover for the preceding financial year if the offender is a company, whichever is higher, where Article 5 has been infringed.

- Up to EUR 15 million, or 3%, whichever is higher, where the infringement concerns obligations of providers pursuant to Art. 16; obligations of representatives pursuant to Art. 25; obligations of importers pursuant to Art. 26; obligations of distributors pursuant to Art. 27; obligations of deployers pursuant to Art. 29; and requirements and obligations of notified bodies pursuant to Art. 33 and Art. 34 paragraphs 1, 3 and 4, and Art. 34bis; and transparency obligations for providers and deployers pursuant to Art. 52.

- 7,500,000, or 1%, where incorrect, incomplete or misleading information is provided, and as above, whichever is higher.

- As an exception in the case of an SME, it is not the higher amount but the lower amount.

With regard to the criteria for graduation, these are the nature, gravity, and duration of the infringement; whether fines have been applied by other market surveillance bodies or by other authorities; the size, turnover and market share of the operator; any other aggravating or mitigating factors, such as profits made or losses avoided; the degree of cooperation with national authorities in order to remedy the infringement and mitigate the damage; the degree of responsibility of the operator in terms of technical aspects and organisational measures implemented; the manner in which knowledge was obtained; intent or negligence; and any measures taken by the operator to mitigate the damage.

On the other hand, it strengthens the fact that Member States must take two actions in this respect, namely to establish other sanctions beyond the fine, and whether or not public administrations are fined.

[17] EUROPEAN COUNCIL OF THE EUROPEAN UNION, Press release "Artificial Intelligence Act: Council and Parliament reach agreement on first rules for AI in the world", 9 December 2023. Updated 2 February 2024.

Artificial Intelligence Law: Council and Parliament reach agreement on world's first AI rules - Consilium (europa.eu)

Finally, and following its passage through the European Parliament in March 2024, the text approved[18] modifies the numbering of this Article 71 to become Article 99. It is also specified in paragraph 1 that the measures established by the States on the system of sanctions and other enforcement measures may be warnings or measures of a non-pecuniary nature.

On the amounts to be imposed as fines, they are the same as those described in the text resulting from the Spanish Presidency, but with a modification of the corresponding number of each Article in relation to fines of up to EUR 15 million or 3% of the infringer's worldwide turnover when infringed: the obligations of providers in Article 16; the obligations of authorised representatives in Article 22; the obligations of importers in Article 23; the obligations of distributors in Article 24; the obligations of deployers in Article 26; the requirements and obligations of notified bodies in Article 31, Article 33(1), (3) and (4) and Article 34; and the transparency obligations of providers and deployers under Article 50.

Also, with some qualification and addition, the criteria for setting fines will be as follows: the nature, gravity, and duration of the infringement and its consequences, taking into account the purpose of the AI system, and, where appropriate, the number of persons affected and the level of damage suffered; whether other supervisory authorities in one or more States have imposed fines on the same operator for the same infringement; whether other authorities have imposed fines on the same operator for other types of infringements arising from the same activity or omission constituting a relevant infringement of the AIA; the size, annual turnover and market share of the offending operator; any other aggravating or mitigating factors such as financial benefits or losses avoided, directly or indirectly through the infringement; degree of cooperation with national authorities in order to remedy the infringement and mitigate its adverse effects; degree of responsibility of the operator, taking into account the technical and organisational measures implemented; how the national authorities became aware of the infringement, in particular whether and to what extent the operator notified the infringement; intentionality or negligence; and the operator's actions to mitigate the harm suffered by the persons affected.

In any case, two aspects are missing. First, a section clearly specifying who may be considered to be the offenders, although it is clear from the

[18] EUROPEAN PARLIAMENT, European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council laying down harmonised rules in the field of Artificial Intelligence (Artificial Intelligence Act) and amending certain legislative acts of the Union.

whole provision that it will be the operators and notified bodies. Secondly, the sanctioning regime seems to be limited to the infringement of the precepts described, so that there may be obligations in the rest of the text which, in the event of non-compliance, would not be sanctioned.

VI. Article 72 on the sanctioning powers of the European Data Protection Supervisor

This second provision completes the sanctioning regime provided for in the GDPR, although its purpose is different, since it mainly regulates the sanctioning powers of the European Data Protection Supervisor (EDPS) with regard to possible breaches by the institutions, agencies and bodies of the European Union to which the rule applies.

We begin, as we did with Article 71, with an analysis of the content of the European Commission's AIA Proposal, the content of which can be divided into three clearly differentiated sections.

The first attributes this sanctioning competence, including the possibility of imposing fines, to the European Data Protection Supervisor, so that no specific body is going to be created within the European Union Administration to exercise supervision in the area of AI. This option could also have been applied in our country, attributing this function to the Spanish Data Protection Agency, although it has fallen to the Spanish Agency for the Supervision of Artificial Intelligence (AESIA). In fact, Article 70 "Designation of competent national authorities", allows it to have been the AEPD, without the need to create a new body.

Furthermore, unlike Article 71, which gives each country the power to decide whether its public sector can be fined, in the case of EU institutions, bodies and agencies, the regulation does provide for this possibility, in line with Regulation 2018/1725, which also gives the EDPS the power to impose fines on the aforementioned when breaches affect the protection of personal data.

In short, it builds on the expertise of the Supervisor who will have a dual supervisory and control role in both Artificial Intelligence and personal data protection. In fact, this body is in favour of the AI supervisory authority being the data protection authority, due to its expertise in managing risks affecting fundamental rights, as well as achieving a consistent application of the standard.[19]

[19] EUROPEAN DATA PROTECTION SUPERVISOR, *Opinion 44/2023 on the proposed*

The second, referring to the most relevant issues concerning the scope of penalties, such as the classification of infringements, fines and criteria for graduation, whose main difference, if we compare it with the content of Article 71, lies in the amount of the fines to be imposed. Thus, they will be considerably lower, being up to 500,000 € for non-compliance with Articles 5 and 11, and up to 250,000 € for the rest of the non-compliances other than the aforementioned precepts.

Cooperation with the EDPS in order to remedy the infringement and mitigate its possible adverse effects, including compliance with his orders, as well as recidivism, on the basis of any previous similar infringement committed by the Union institution, agency or body, are also introduced as new criteria for the graduation of the amounts. Strangely, these criteria are not among those foreseen in Article 71, as they are also applicable when a market surveillance authority is acting. The second criterion does appear to some extent, but it is worded in a more complex way so that we cannot say that its content qualifies as "recidivism". The other criteria contemplated in Article 72 are similar to those already mentioned in Article 71, such as graduation, which, in addition to the nature, seriousness and duration of the infringement, must be assessed.

As for the third, it complements the previous one by regulating elements of the sanctioning procedure itself, such as the right to be heard that the alleged offender may exercise before being sanctioned and the right of access to the file, without prejudice to guaranteeing the legitimate interest of individuals and companies in protecting their personal data or commercial secrets.

For its part, the text of Article 72 of the Council of the European Union's proposal for the AIA of 6 December 2022 does not propose any changes to its content.

On the other hand, there are changes in the text of the European Parliament's amendments, which can also be classified into two groups: one on the criteria for the quantification of fines; and the other on the classification and amount of fines, which, as we shall see, is where the greatest changes are to be found.

Regarding the first, new criteria are introduced, many of which have been proposed for inclusion in Article 71, a logical question since it would not make much sense for some to be applied in the scope of application of this provision and others in that of Article 72, without prejudice to the fact that there may be some that are not applicable if we take into account the nature of the bodies and institutions of the European Union. Thus, it is proposed

Artificial Intelligence Act in the light of legislative developments, 2023. https://edps.europa.eu/system/files/2023-10/2023-0137_d3269_opinion_en.pdf to include in addition to the nature, gravity, and duration also the purpose of the AI system, the number of persons, harm suffered and any previous infringement; any measures to mitigate the harm to persons (although this criterion appears to some extent already in the text of the Commission's draft AI); the degree of responsibility of the Union institution, agency or body, considering the technical and organisational measures it applies; the way in which the EDPS has become aware of the infringement; and the annual budget of the body. In addition, it should be added that the fines will not affect the effective functioning of the sanctioned Union institution, body, office or agency, which can also be used as a criterion for the imposition of the fine.

On the second, non-compliance with the prohibitions of Article 5 will be fined up to 1.5 million euros; Article 10 up to 1 million euros; and the rest 750,000 euros, which shows that the amounts of the fines are considerably higher compared to the text of the Commission's draft AIA.

With respect to the text that emerged from the Spanish Presidency and the trialogues cited above, as with Article 71, there are modifications in terms of the amounts and infringements, with two groups. The infringement of Article 5 can reach up to €1,500,000, and any other different infringement affecting other provisions of the regulation, up to €750,000.

As for the final approved text of the AIA, this Article 72 is renumbered as Article 100, the amounts are identical to the previous ones, as well as the graduation criteria. The greatest novelty appears in the following article, Article 101, which grants the Commission the power to fine providers of general-purpose AI models that do not exceed 3% of worldwide turnover or 15 million euros, whichever is higher, when any of the following conducts take place, either through fault or negligence: failure to comply with the AIA; failure to comply with the request for information in Article 91, or providing it in an incorrect, incomplete, or misleading manner; failing to comply with a requested Article 93 measure; or failing to give the Commission access to the general purpose or systemic risk AI model for the purpose of carrying out the Article 92 assessment.

In other words, in addition to the sanctioning power exercised by the States, the Commission itself may also impose fines, provided that the requirements described above are met.

VII. Conclusions

The main premise of the regulation of the sanctioning regime, regardless of the matter in question, is to provide legal certainty so that the addressees of the rule can know at all times the facts for which they may be sanctioned, the amounts, and up to when. From the content analysed in its different versions, three issues emerge that are not fully regulated in such a way that legal certainty cannot be achieved. We are referring to the absence of alleged offenders, the deficient technique for defining offences, and the failure to regulate the statute of limitations. Curiously, these three deficiencies also appear in the GDPR, so the same errors are repeated, although, with regard to the part on personal data protection, some of them have been solved to a certain extent.

With regard to the parties presumed responsible for committing an infringement, although it is clear from Article 71 itself that it will apply to all operators, it would have been clearer to have a section establishing this in a more specific manner.

With regard to infringements, although it is true that a series of precepts have finally been delimited on whose non-compliance the sanctioning regime would apply, such as Articles 5, 16, 22, 23, 23, 24, 26, 31, 33 sections 1, 3 and 4, 34 and 50, which means, as we explained above, going precept by precept to assess where an infringement can be committed and where it cannot. For this reason, we consider that there should be an Annex describing the possible infringements, indicating the punishable conducts, which could be completed by also determining the statute of prescriptions.

Otherwise, and given that such legal certainty is still necessary, it will be our legislator who will try to provide it, as he has done with the LOPDGDD, albeit bearing in mind that he cannot typify the offences, given that such typification, albeit in a very broad form, is in the AIA. It is likely to adopt the same solution as in the aforementioned LOPDGDD: description of punishable conduct for the purposes of setting the prescription period.

On the other hand, it is rather striking that the only sanctioning measure envisaged is a fine, when other instruments can also be used. We therefore consider that warnings, cautions, enforcement measures, as well as the withdrawal, prohibition, and restriction of an AI product should be included. The last three, on the other hand, do appear in Article 65.

There is also a need to complete the criteria for imposing fines, with the European Parliament proposing the most amendments to introduce new criteria. It should be recalled that the Commission's text only included three.

In short, the classification of offences and criteria for the graduation of fines should be improved, as well as the introduction of those who may be presumed responsible, the statute of limitations, and more measures to be applied other than just a fine.

As a complement to all this, and taking into account the experience of the GDPR, we consider that the regulation should not leave it up to each country to decide whether or not to apply fines to its public sector, since, if so, some countries, probably the majority, will consider it, but some will not, losing uniformity. This possibility of fining should be included. Moreover, there is a certain contradiction in the fact that the regulation leaves it open but grants this power to the EDPS with regard to the bodies and institutions of the European Union.

Finally, and to conclude, we applaud the fact that the EDPS has been attributed this sanctioning power in relation to the aforementioned, without the need to create a specific body, which results in greater effectiveness, efficiency, and time reduction. If there were two bodies, one for data protection and the other for AI, as is the case in our country, close cooperation between them would be indispensable, especially in view of the implications between the two areas.

RIGHT TO LODGE A COMPLAINT AND RIGHT TO AN EXPLANATION. MEANS OF REDRESS FOR INDIVIDUALS IN THE ARTIFICIAL INTELLIGENCE ACT

Aurelio Lopez-Tarruella Martinez
Senior Lecturer in International Private Law
University of Alicante

I. Introduction

The purpose of this paper is to analyse Section 4 ("Remedies") of Chapter IX ("Post-market surveillance, information exchange, market surveillance") of the AIA, which covers Articles 85 to 87. As its title indicates, this section sets out the remedies available to individuals against non-compliance with the Regulation by providers, deployers or any other operator involved in the AI value chain.

This is a Section introduced by the European Parliament which did not appear in the European Commission's initial Proposal or in the Council's Common Position. It strengthens the position of persons who may be affected by decisions taken on the basis of information provided by AI systems. This was an omission that had been criticised by the doctrine in the Commission's initial proposal. In fact, this section is the only one in the Regulation that includes rights for persons affected by the operation of an AI system: the right to lodge a complaint; and the right to an explanation of decisions taken individually. In any case, as we shall see, the text finally adopted is not entirely satisfactory in that it reduces the effectiveness of these rights.

The paper is divided into three parts. The first two parts deal respectively with Article 85, which regulates the right to lodge a complaint with a market surveillance authority; and Article 86, which deals with the right to an explanation of decisions taken on an individual basis. The third part explains Articles 87 and 110, which have an auxiliary nature and contain, respectively, a reference to Directive 2019/1937 on the protection of persons who report breaches of Union Law[1]; and an amendment to the Annex to Directive 2020/1828[2] on representative actions for the protection of the collective in-

[1] Directive (EU) 2019/1937 of the European Parliament and of the Council of 23 October 2019 on the protection of persons reporting breaches of Union law, available at http://data.europa.eu/eli/dir/2019/1937/oj.

[2] Directive (EU) 2020/1828 of the European Parliament and of the Council of 25 November 2020 on representative actions for the protection of the collective interests of

terests of consumers, ensuring that associations can bring collective actions for non-compliance with the AI Act.

II. The right to lodge a complaint with a market surveillance authority

The doctrine has rightly held that European digital laws[3], including the AIA, are inspired by the General Data Protection Regulation 2016/679 (hereinafter GDPR) both in their structure and in the content of some of their provisions.[4]

At first sight, it could be assumed that Article 85 is a manifestation of this inspiration in that its title and the content of the first paragraph is similar to the regulation of the right to lodge a complaint contained in Article 77 GDPR, Article 53 DSA, Article 14.1 *in fine* and 24.1 *in fine* DGR, and Article 38 RD. This interpretation is supported by the wording of Article 110, which opens the way for, as in these instruments, bodies for the collective representation of consumers' interests to lodge, within the framework of Directive 2020/1828, complaints for non-compliance with the Regulation.

However, this assumption is discredited by the reference that paragraph 2 of Article 85 makes to Regulation 2019/1020 on market surveillance[5] for the purposes of regulating these claims. This is because this reference entails the attribution to this right of a completely different content to that established in the GDPR and the rest of the digital laws. In my opinion, this special regulation greatly reduces the usefulness and practical effectiveness of this right, a circumstance that may constitute an obstacle to achieving the objectives of the Regulation.

consumers and repealing Directive 2009/22/EC, available at http://data.europa.eu/eli/dir/2020/1828/oj.

[3] For the purposes of this paper, in addition to the AIA, 'European digital laws' are: Data Governance Regulation 2022/868 (DGR), Digital Markets Regulation 2022/1925 (DMLR); Digital Services Act 2022/2065 (DSA) and Regulation 2023/2854 on harmonised rules for fair access to and use of data (RD).

[4] In the same vein, Gascón Macén, A., "El Reglamento General de Protección de Datos como modelo de las recientes propuestas de legislación digital europea", *CDT*, Vol 13(2), 2021, pp. 209-232. https://doi.org/10.20318/cdt.2021.6256; Papakonstantinou, V. / De Hert, P. "Post GDPR EU laws and their GDPR mimesis. DGA, DSA, DMA and the EU regulation of AI", *European Law Blog*, 1 April 2021, https://europeanlawblog.eu/2021/04/01/post-gdpr-eu-laws-and-their-gdpr-mimesis-dga-dsa-dma-and-the-eu-regulation-of-ai/.

[5] Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and product conformity and amending Directive 2004/42/EC and Regulations (EC) 765/2008 and (EU) 305/2011, available at http://data.europa.eu/eli/reg/2019/1020/oj.

According to this provision, the competence to hear (or rather "take into account") these complaints does not correspond to a single authority (in the case of Spain, to AESIA), but to several, depending on the classification of the AI system. This circumstance may make the exercise of this right more difficult and further undermine its practical usefulness. Moreover, the margin of discretion established by this provision may generate undesired situations of *forum shopping* between the market surveillance authorities of different Member States.

1. Developments in the text of the provision in the preparatory work

As indicated in the Introduction, the initial proposal of the European Commission did not include this right, a circumstance criticised by the doctrine[6] and in the Joint Opinion of the European Data Protection Board and the European Data Protection Supervisor[7]. Despite this, the Council did not see fit to include this right in its Common Position of November 2022[8]. The European Parliament did so in Amendments 628 and 629 to the Commission Proposal adopted on 14 June 2023[9]. The inclusion of this right was accompanied by other complementary provisions also inspired by the GDPR:

> *Article 68a. Right to lodge a complaint with a national supervisory authority*
>
> *1. Without prejudice to any other administrative remedy or judicial proceedings, any natural person or group of natural persons shall have the right to lodge a complaint with a national*

[6] Smuha, N. et al, "How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act", 2021, available at https://ssrn.com/abstract=3899991; Ebers, M. et al, "The European Commission's Proposal for an Artificial Intelligence Act-A Critical Assessment by Members of the Robotics and AI Law Society (RIALS)", *J*, vol. 4, 2021, pp. 589-603. https://doi.org/10.3390/j4040043; Veale, M. and Zuiderveen Borgesius, F.J., "Demystifying the Draft EU Artificial Intelligence Act - Analysing the good, the bad, and the unclear elements of the proposed approach", *Computer Law Review International*, vol. 22, 2021, pp. 97-112, esp. 111; Lúcia Raposo, V. "Ex machina: preliminary critical assessment of the European Draft Act on Artificial Intelligence", *International Journal of Law and Information Technology*, Vol. 30, Issue 1, 2022, p. 102.

[7] EUROPEAN DATA PROTECTION BOARD / EUROPEAN DATA PROTECTION SUPERVISOR, "Joint Opinion 5/2021 on the Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)", 18 June 2021, p. 18, available at https://edpb.europa.eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-52021-proposal_en.

[8] Council Common Position of 22 November 2022 (available at https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf).

[9] https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_ES.html

> *supervisory authority, in particular in the Member State in which he has his habitual residence, place of work or place of the alleged infringement, if he considers that the AI system concerning him infringes this Regulation.*
>
> *2. The national supervisory authority with which the complaint has been lodged shall inform the complainant of the progress and outcome of the complaint, including the possibility of access to judicial protection under Article 78.*
>
> *Article 68b. Right to an effective remedy against a national supervisory authority*
>
> *1. Without prejudice to any other administrative or non-judicial remedy, any natural or legal person shall have the right to effective judicial protection against a legally binding decision of a national supervisory authority concerning that person.*
>
> *2. Without prejudice to any other administrative or extra-judicial remedy, any natural or legal person shall have the right to an effective judicial remedy where the competent national supervisory authority pursuant to Article 59 fails to act on a complaint or fails to inform the person concerned within three months of the action taken or the outcome of the complaint lodged pursuant to Article 68a.*
>
> *3. Actions against a national supervisory authority shall be brought before the courts of the Member State in which the national supervisory authority is established.*
>
> *4. Where an action is brought against a decision of a national supervisory authority which has been preceded by an opinion or a decision of the Commission in the framework of the Union safeguard procedure, the supervisory authority shall refer that opinion or decision to the court.*

Article 68a proposed by the Parliament is an almost verbatim copy of Article 77 GDPR. There is no doubt that this is due to the mistake of referring in paragraph 2 to "the possibility of access to judicial protection under *Article 78*". This provision, *but of the GDPR*, is the one that regulates the "right to effective judicial protection against a supervisory authority". The reference should be to Article 68b, the provision that regulates this possibility in the AIA. The error is demonstrative of Parliament's desire to introduce a right to lodge a complaint with the same content as Article 77 GDPR and other European digital laws.

The need to reach a compromise text at the trialogue stage entails two important changes. On the one hand, Article 68b ends up having a substantially different wording from the one proposed by the Parliament. These differences are particularly important with regard to the current paragraph 2 of Article 85.

> *Article 85. Right to lodge a complaint with a market surveillance authority*
>
> *Without prejudice to other administrative or judicial remedies, any natural or legal person having grounds to consider that there has been an infringement of the provisions of this Regulation may submit complaints to the relevant market surveillance authority.*
>
> *In accordance with Regulation (EU) 2019/1020, such complaints shall be taken into*

account for the purpose of conducting market surveillance activities, and shall be handled in line with the dedicated procedures established therefor by the market surveillance authorities.

On the other hand, Article 68b of the Parliament's proposal is deleted. This is because, as will be explained in the following section, according to Regulation (EU) 2019/1020, the market surveillance authority is not obliged to adopt a decision on the complaint. Consequently, the right to lodge an appeal against such a decision before a judicial authority becomes meaningless. This shows that the regulation of this right in the AIA is completely different from that provided for in the GDPR and the other European digital laws.

The main reason that may explain the final wording of Article 85 is that, unlike the GDPR and other digital laws, the AIA does not attribute the function of market surveillance of high-risk AI systems to a single authority but to several. In addition to AESIA, the AIA assigns powers to the authorities designated by the instruments implementing the regulations set out in Annex I, Section A (instruments that form part of the new regulatory framework setting market entry requirements on machinery, safety of toys, medical devices, aviation vehicles, etc.). These authorities, whose activities are regulated by Regulation (EU) 2019/1020, do not have among their functions to handle individual complaints, especially considering the complexities surrounding the ones concerning an AI system's AIA compliance.

Undoubtedly, in line with the European Parliament's amendments, the choice could have been made to confer competence to hear such complaints exclusively on one authority (AESIA). In fact, Article 74.3 allows Member States to confer, "in appropriate circumstances", competence on a market surveillance authority other than the one designated by the legal acts of the new regulatory framework. Thus, for example, the competence to deal with all complaints from individuals could be concentrated in the AESIA. However, this could give rise to two types of problems: uncoordinated and overlapping functions between AESIA and market surveillance authorities in specific sectors; and *forum shopping* in relation to authorities in other Member States, as will be explained below.

This being the main reason, the question remains as to whether the emptying of the right of complaint is part of the trend observed at the end of the negotiations to lower the requirements and obligations of certain categories of AI systems in order to facilitate their development in the European Union.

It could also be argued that the absence of a genuine right to lodge a complaint would be compensated by the obligation for those responsible for deploying high-risk AI systems, introduced in the latest version of the Regulation, to carry out fundamental rights impact assessments (Article 27).

In any case, even if the final regulation can be justified on the above grounds, this does not prevent us from stating, as we explain below, that the achievement of the objectives of the AIA (Article 1) may be hampered.

2. Differences between the regulation of the right to lodge a complaint in the AI Act and that established in the GDPR and other European digital laws

The doctrine considers the right to lodge a complaint contained in the GDPR to be a very useful tool for two reasons. Firstly, because, as the doctrine[10] and the CJEU[11] have shown, it favours the defence of the fundamental right to the protection of personal data, by offering data subjects a simpler and free way to assert their rights. Secondly, it gives controllers and processors an incentive to comply with the Regulation, since, if supervision is left to public authorities alone, it may be affected by reasons relating to the lack of resources of these authorities, the limited expertise of their staff, or political or geostrategic reasons (think of Ireland or Luxembourg). The introduction of this right allows individuals to participate in market surveillance and to claim their rights, which gives companies an incentive to comply with their regulatory obligations[12]. It is clear that the actions carried out by Maximilian Schrems, *NOYB* or *La Quadrature du Net* have given a great boost to effective compliance with the GDPR.

The usefulness of this right can be transferred to all those regulations in which the right has been included with a regulation similar to that provided for in the GDPR. This is not the case of the AIA because, as we have said, the regulation is different.

In the GDPR, this right has an unwaivable minimum content established in the regulation itself and in the case law of the CJEU. Thus, Article 77.2 GDPR states that "[t]he supervisory authority with which the complaint has been lodged shall inform the complainant on the progress and the outcome

[10] De Miguel Asensio, P., *Derecho privado de Internet*, 6th Ed, Civitas, Madrid, 2022, p. 537; AGENCY FOR FUNDAMENTAL RIGHTS OF THE EUROPEAN UNION, COUNCIL OF EUROPE, EUROPEAN DATA PROTECTION SUPERVISOR, EUROPEAN COURT OF HUMAN RIGHTS, *Manual of European Data Protection Law. 2018 Edition*, Publications Office of the European Union, 2019, https://data.europa.eu/doi/10.2811/60145

[11] C-203/15, "Tele2 Sverige", nr. 123; 6 October 2015, "Schrems", nr. 41; 8 April 2014, "Digital Ireland", nr. 68.

[12] Of course, the success of the system depends on Member States providing these authorities with the necessary resources to carry out their work effectively.

of the complaint, including the possibility of a judicial remedy pursuant to Article 78 (right to an effective judicial remedy)". Furthermore, the CJEU in its judgment of 7 December 2023, C-26/22, *SCHUFA*, has established that the provision obliges the supervisory authority to adopt an administrative decision that is subject to full judicial review, relating to substantive arguments and not exclusively to procedural issues[13]. The complaint procedure is not similar to a petition, but is conceived as a mechanism capable of effectively protecting the rights and interests of the data subjects.[14]

With regard to those questions on the complaint not expressly regulated in the Regulation, the CJEU of 12 January 2023, C-132/21, *Budapesti Elektromos Művek*, recalls that it is up to the Member States, in application of the principle of procedural autonomy, to determine the procedures on the basis of which this right must be articulated[15]. However, the regulation of these procedural channels must not jeopardise the useful effect and effective protection of the right to lodge a complaint[16]. Similarly, "such regulation must not be less favourable than that concerning similar remedies established for the protection of rights recognised by the internal legal order (principle of equivalence) or make it impossible in practice or excessively difficult to exercise the rights conferred by the legal order of the Union (principle of effectiveness)"[17].

Finally, it follows from the case law of the CJEU that the interpretation of these provisions must take into account Recitals 10 and 11 of the Act. According to Recital 10, the objective of the Regulation is to ensure a high level of protection for natural persons with regard to the processing of personal data in the Union. Recital 11 of the same Regulation furthermore states that effective protection of these data requires that the rights of data subjects are strengthened.[18]

[13] Paragraph 70.

[14] Paragraph 58.

[15] Paragraph 45: "In the absence of Union legislation in this area, each Member State must, in accordance with the principle of the procedural autonomy of the Member States, lay down rules governing administrative and judicial procedures designed to ensure a high level of protection of the rights conferred on individuals by Union law".

[16] Paragraph 47: "[…] the regulation of the application of such concurrent and independent remedies must not jeopardise the effectiveness and effective protection of the rights guaranteed by that Regulation".

[17] The existence of this case law has not been able to prevent the emergence of significant differences in the domestic regulations of the right of complaint provided for in Art. 77. See EDPS, *Study on the National Administrative Rules Impacting the Cooperation Duties for the National Supervisory Authorities - Final Report*, EDPS 2019/02-07, 2020, available at https://edpb.europa.eu/system/files/2023-04/call_7_final_report_07012021.pdf.

[18] App. 61, CJEU of 7 December 2023, C-26/22, *SCHUFA*.

This case law is difficult to transfer to the interpretation of Article 85 AIA. In this case, according to paragraph 2, the regulation of the processing and effects of the complaint is referred to Regulation 2019/1020. That regulation does not regulate a right of complaint within the meaning of the GDPR. Article 11.3 of that Regulation, the only one which contains a reference to consumers and economic operators, indicates that consumers and economic operators may lodge a complaint which the supervisory authority will take into account, among other factors, in deciding what checks to carry out for the purposes of determining whether the Regulation is complied with:

> *"In deciding which checks to carry out, on which types of products and on what scale, market surveillance authorities will follow a risk-based approach, taking into account the following factors:*
>
> *(e) consumer complaints and other information received from other authorities, economic operators, the media and other sources that may indicate non-compliance".*

There is no obligation in this provision, or in any other provision of the Regulation, for the market surveillance authority to inform the complainant about the course and outcome of the complaint. Nor is there an obligation as laid down by the CJEU in relation to the GDPR to provide a substantive response to the complaint. The wording of Article 85 is explained by the need to align it with the regulation in Regulation 2019/1020: "*complaints shall be taken into account for the purposes of conducting market surveillance activities, and shall be handled in line with the dedicated procedures"*. It can be inferred from this provision that Member States are free to decide on the specific procedures for handling such complaints, and the value that can be attributed to them. All that Article 85(2) AIA states and Article 11.3 Regulation 2019/1020 reiterates is that such complaints "shall be taken into account". But there is no obligation for supervisory authorities to deal with such complaints on an individual basis, or to explain the reasons why such complaints, if any, are not finally taken into consideration when carrying out checks or initiating investigations. In short, there is no obligation on the market surveillance authority to take a decision on the complaint.

The particular nature that Article 85 attributes to the right to lodge a complaint explains that, unlike in the GDPR (and other digital laws), individuals cannot appeal the decision or the omission of a decision by the market surveillance authority before the contentious-administrative jurisdiction. As a result, the article proposed by the Parliament (Article 68b), which included this right, has been deleted from the final text.

The scant regulation of this right in Article 85 AIA also leaves a number of questions. To begin with, it remains to be seen whether our legislator will adopt a special regulation in relation to complaints concerning AI systems whose oversight is the responsibility of AESIA. In principle, the reference in Article 74 to Regulation 2019/1020 applies to all "AI systems covered by the […] Regulation", which suggests that complaints to AESIA could be treated in the same way as in that treatment. However, there is nothing to prevent the adoption of a regulation similar to that established in the GDPR in which, at least, AESIA is obliged to respond to the complainant, and that the decision may be appealed before the courts of the contentious-administrative order. This would undoubtedly help to protect the objectives of the Regulation, but it has two disadvantages. First, it would create a comparative disadvantage: while users of AI systems subject to AESIA oversight would have a real right of complaint, the rest would not. Second, if the other Member States did not adopt a regulation similar to the one proposed, *forum shopping* could arise: users (and, in particular, the associations that represent them) would choose to bring complaints before the authorities of Member States with a regulation that is more beneficial to their interests in terms of the right to lodge a complaint.

The reference in Article 85(2) to Regulation 2019/1020 raises doubts in relation to high-risk AI systems in the financial sector, and those used for law enforcement, border management, administration of justice and democratic processes. As we will see below, according to Article 74.6 and 8, the competent authority in these cases is the National Securities Market Commission and the AEPD respectively.

It seems logical to think that complaints to the first of these authorities will be processed in accordance with the specific regulations of the financial sector, so that the reference in Article 85.2 to Regulation 2019/1020 is meaningless. This clarification is relevant because the National Securities Market Commission (CNM) has its own complaints service[19]. However, it should be recalled that Article 74.7 allows Member States, in appropriate circumstances and provided that coordination is ensured, to confer the competence to supervise the application of the AIA to a different authority (which could be AESIA).

It is more difficult to reach a conclusion in the case of the AEPD because, strictly speaking, the complaint filed on the basis of Article 85 AIA will not be for a breach of the GDPR. But neither is it logical for the AEPD to end up applying a Regulation (2019/1020) that does not fall within its powers.

Finally, in view of the fact that Article 85.1 AIA does not provide an

[19] https://www.cnmv.es/portal/inversor/reclamaciones.aspx

effective mechanism for individuals to complain about a breach of the Regulation, it is worth considering whether there is another way of doing so. In this regard, one could consider bringing a civil court action against the allegedly non-compliant operator seeking the cessation of an activity that does not comply with the Regulation or compensation for the damage that such activity may have caused. I do not believe that the fact that, unlike Article 79 GDPR, the text of the Regulation does not expressly provide for this possibility is an impediment. Moreover, it should be recalled that Article 85.1 starts by stating that the right to lodge complaints is enjoyed "*[w]ithout prejudice to other administrative or judicial remedies*".

The problem with this route is that it involves high costs and the procedure can take a long period of time. This is, in my view, a viable option only for collective representation bodies. It certainly remains to be seen what happens in practice, but the doctrine has raised this scenario in relation to the other European digital law where the right to lodge a complaint is also not regulated: the Digital Markets Act.[20]

3. Regulation of the right to lodge a complaint in the Act

Having explained what it would have been desirable for the right to lodge a complaint under the AIA to be, but is not, it then proceeds to explain what it ultimately is. For these purposes, the right can be understood as a mere power of any natural or legal person to lodge a complaint with market surveillance authorities reporting an alleged breach of the AIA by an operator covered by the Regulation. Such a complaint may also relate to the failure of the deployer to comply with the obligation laid down in Article 86 to provide an explanation of the decision taken on the basis of the results provided by certain high-risk AI systems.

As explained above, according to Article 85.2 AIA and Article 13.1 Regulation 2019/1020, the authorities' only obligation is to take such complaints into account when deciding whether to initiate checks which, where appropriate, will give rise to investigations within the meaning of Articles 79 and 80 AIA. An exception to this general rule is made for the CNMV, which, in our view, will have to deal with complaints relating to AI systems used in the financial sector in accordance with its particular rules.

Having explained these issues in the previous section, it is appropriate in

[20] G. Monti "Procedures and institutions" in AAVV, *Effective and Proportionate Implementation of the WFD*, CERRE, 2023 pp. 164 ff, esp. 181, available at https://cerre.eu/wp-content/uploads/2023/01/DMA_Book-1.pdf

the following lines to clarify some other issues relating to this right to lodge a complaint.

First, Article 85 indicates that the standing to lodge a complaint lies with "any natural or legal person having grounds to consider that there has been an infringement of the provisions of this Regulation". As we will see in Section IV, according to Article 87, these persons can benefit from the protection offered by Directive 2019/1937 on the protection of persons who report breaches of Union law.

Unlike Article 80 GDPR (and other European digital laws), the AIA does not expressly regulate the possibility for individuals affected by a breach to mandate a non-profit entity, organisation or association to represent them before the competent authority. This lack of regulation is undoubtedly due to the completely different meaning of the right of complaint in the AIA. However, such representation seems possible for two reasons. First, because otherwise the mandate of Article 110 to include the AIA in the Annex to Directive 2020/1828 on the protection of the collective interests of consumers would be meaningless. Second, because Article 9 of Regulation 2019/1020 provides for the power of market surveillance authorities to agree with "organisations representing economic operators or end-users" on joint activities with a view to encouraging compliance or detecting cases of non-compliance.

Secondly, the question of determining the authority with which the complaint should be lodged arises from a twofold dimension: the material dimension (determination of the authority with jurisdiction over the matter) and the territorial dimension (determination of the Member State before whose authority the complaint must be made).

In relation to the first dimension, Article 85 should be read in conjunction with a number of provisions determining the competent authorities to carry out market surveillance of AI systems. In this respect, the following classification should be made.

a) AI systems based on a general purpose AI model. The competence lies with the AI Office, which "shall have all the powers of a market surveillance authority within the meaning of Regulation 2019/1020" (Article 75).

(b) High-risk AI systems intended to be used as a component part of a product or constituting in itself a product covered by the harmonisation legislation listed in Annex I, Section A (Article 74.3). As will be recalled, this legislation, which is part of the "New Regulatory Framework", covers, inter alia, toys, machinery, boats, lifts, radio equipment, medical devices or motor vehicles. In these cases, the complaint must be submitted to the authority designated in each legislative instrument. Thus, in Spain, depending on the product, the competence may correspond to agencies or administrative units belonging to multiple ministries.[21]

c) High-risk AI systems marketed, put into service or used by financial institutions regulated by the relevant Union legislation (Article 74.6). In these cases, the competent authority to hear the complaint is the National Securities Market Commission (CNMV). However, as mentioned above, this attribution of competence could change, as paragraph 7 states that "in appropriate circumstances and provided that coordination is ensured, another relevant authority may be identified by the Member State as market surveillance authority for the purposes of this Regulation".

(d) Systems are used for law enforcement purposes and for the purposes listed in points 6, 7 and 8 of Annex III, i.e., law enforcement matters; migration management, asylum and border control; and the administration of justice and democratic processes (Article 74.8). In this case, the competence to hear the complaint lies with the AEPD.

(e) Prohibited AI practices, high-risk AI systems not listed in Annex III (stand-alone AI systems) and non-high-risk AI systems. The competence to hear complaints in these cases lies with AESIA.

As indicated above, the distribution of competence to deal with complaints among a plurality of authorities may hinder the exercise of this right. To reduce this problem, the obligations foreseen in Article 70.2 for Member States are of great relevance: designation of a single contact point; and making available to the public, by electronic means of communication, information on how to contact the competent authorities and the single contact points. In addition, the European Commission is obliged to publish online a list of such contact points at European level.

In relation to the territorial dimension, the question arises as to before which Member State's authorities should the claim be made? In the context of the GDPR, this question is of particular relevance because of the nature of the complaint and the need to facilitate its submission by individuals. This leads the European legislator to establish special rules of jurisdiction that attribute competence to the authorities of the residence of the data subject, which is an exception to the general rule of the Regulation according to which the competence to supervise controllers and processors lies with the authorities of the Member States of establishment (or, as the case may be, of the main establishment).[22]

[21] The list of market surveillance authorities can be found at https://single-market-economy.ec.europa.eu/single-market/goods/building-blocks/market-surveillance/organisation_en.

[22] Art. 56 GDPR.

In the case of the right to lodge an Article 85 AIA complaint, this attribution of competence is of lesser importance for two reasons. First, the particular nature of the complaint, which, as has been argued, will be dealt with in accordance with Regulation 2019/1012 and authorities may or may not take into account. Second, because the AIA has opted for a decentralised system of jurisdiction: market surveillance authorities have jurisdiction to hear complaints relating to infringements occurring in the territory of their Member State[23]. In those cases (presumably common in practice) where the infringement of the AI system occurs in more than one Member State, this decentralised system of jurisdiction allows the complaint to be brought before the authority of any of those States. This opens the door to *forum shopping*: it is to be expected that individuals, and in particular collective representative associations, will choose to file a complaint in Member States that provide for more user-friendly procedural routes for filing, or that are more likely to be more likely to initiate investigations, or that simply have greater resources, or greater expertise and know-how. Hence, as mentioned above, the concentration of competence to hear individual complaints in the EIOPA, although beneficial for individuals, may have a dangerous "pull effect".

III. The right to an explanation of individual decision-making

The second right granted to individuals in the AIA is the right to obtain an explanation of individual decision-making. This reinforces the requirement of explainability, in accordance with the European commitment to reliable AI, that all AI systems developed and marketed in the Union must have. This is a right inspired by Article 22 of the GDPR, which is why we consider it appropriate to refer to the *Guidelines on automated individual decisions* of the former Article 29 Working Party[24]. It is precisely the relationship with this provision of the GDPR that is the main question posed by Article 86, which is why we will devote a specific section to it. It is also relevant to analyse the relationship of this right with the protection that intellectual property rights and trade secrets can provide to many of the elements present in an AI system.

[23] A comparative analysis of the advantages and disadvantages of different supervisory systems for European regulatory instruments can be found in Monti, G., De Streel, A., "Improving EU Institutional Design to Better Supervise Digital Platforms", *CERRE Report*, 2022, available at https://cerre.eu/publications/improving-eu-institutional-design/.

[24] ART. 29 WORKING GROUP, "Guidelines on automated individual decisions and profiling for the purposes of Regulation 2016/679 of 3 October 2017", available at https://ec.europa.eu/newsroom/article29/items/612053.

1. Developments in the text of the provision in the preparatory work

The right to an explanation has followed a similar path to the right to lodge a complaint under Article 85 AIA. Neither the Commission's initial Proposal nor the Council's Common Position of December 2022 contained any reference to it. The first reference to it is to be found in Amendment 630 of the European Parliament Report of June 2023, where it is proposed to introduce a new Article 68c with the following wording:

> *1. Any affected person subject to a decision taken by the deployer on the basis of output information from a high-risk AI system which produces legal effects or significantly affects him in a way that he considers prejudicial to his health, safety, fundamental rights, socio-economic well-being or any other of his rights deriving from the obligations laid down in this Regulation, shall have the right to request from the deployer a clear and meaningful explanation, in accordance with Article 13.1, of the role of the AI system in the decision-making procedure, the main parameters of the decision taken and the relevant input data.*
>
> *2. Paragraph 1 shall not apply to the use of AI systems for which national or Union law provides for exceptions or restrictions to the obligation laid down in paragraph 1 in so far as such exceptions or restrictions respect the essence of fundamental rights and freedoms and are a necessary and proportionate measure in a democratic society.*
>
> *3. This Article shall apply without prejudice to Articles 13, 14, 15 and 22 of Regulation (EU) 2016/679.*

The text finally adopted does not vary much, although it includes important clarifications:

Article 86 Right to an explanation of individual decision-making

> *1. Any affected person subject to a decision which is taken by the deployer on the basis of the output from a high-risk AI system listed in Annex III, with the exception of systems listed under point 2 thereof, and which produces legal effects or similarly significantly affects that person in a way that they consider to have an adverse impact on their health, safety or fundamental rights shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken.*
>
> *2. Paragraph 1 shall not apply to the use of AI systems for which exceptions from, or restrictions to, the obligation under that paragraph follow from Union or national law in compliance with Union law*
>
> *3. This Article shall apply only to the extent that the right referred to in paragraph 1 is not otherwise provided for under Union law.*

The changes between the two texts that are worth noting are as follows:

a) The final text limits the scope of application of the right to high-risk AI schemes in Annex III, with the exception of those provided for in point 2.

b) The wording of the first paragraph is simplified by reducing the legal interests that may be affected by the decision to "health, safety and fundamental rights", which is appropriate in that it is in line with the objectives pursued by the AIA (see Article 1) and avoids the ambiguity which, in my view, could be caused by the other legal interests mentioned in the initial version ("*socio-economic, well-being or any other of their rights deriving from the obligations laid down in this Regulation").*

c) The obligation on the deploying officer is relaxed by referring more generically to the "decision-making procedure and the main elements of the decision taken". The explicit reference in the Parliament's version of the report to "the main parameters" and "the relevant input data" could make it difficult to comply with the obligation and also raise doubts as to whether they were obliged to disclose proprietary or trade secret information.

d) From the exclusion in paragraph 2, the condition that exceptions or restrictions "respect the essence of fundamental rights and freedoms and are a necessary and proportionate measure in a democratic society" is removed. The exclusion is appropriate in that it appears to be a condition which could give rise to interpretative problems, and which ignores the fact that, by definition, EU or national rules which may provide for such exceptions should, by definition, meet these requirements because they are rules where the rule of law and European values are respected.

e) The exclusion in paragraph 3 replaces the explicit reference to the Articles of the GDPR (*Articles 13, 14, 15 and 22)* with a generic reference to "Union law". In any event, as will be seen throughout the analysis, the GDPR is the main one affected by the provision.

2. The Explainability Principle for Artificial Intelligence Systems

The use of complex algorithms to automate decision-making in our day-to-day lives has become commonplace. It is present in personnel selection processes in public or private companies, the granting of credit, the price of an insurance policy, or the personalisation of advertising received by the user of a social network. In general, the functioning of these algorithms is unknown and unintelligible not only to the people affected by the automated decisions, but also to the company using the AI system. This is particularly the case if the algorithm is based on machine learning and, more specifically, deep learning techniques. This has led the doctrine to coin the term "black box" society, referring to the fact that the decisions that drive our daily lives, which are becoming increasingly relevant and far-reaching, are taken by systems that we do not understand and therefore cannot know what they are based on[25].

The opacity of AI systems is a problem not only because it is an obstacle to everyone's right to know the reasons behind decisions that affect them, but also because this lack of knowledge makes it impossible to investigate the reasons why an algorithmic system makes mistakes. These errors can have material consequences (such as, for example, the malfunctioning of a device connected to the Internet of Things); and also personal consequences, which can range from a physical injury (resulting, for example, from an accident caused by an autonomous vehicle[26]) to economic or moral damage resulting from being rejected in a recruitment process, being denied a loan by a bank, not being accepted to participate in public competitions, or being classified as a person with a high risk of absconding[27]. These errors can lead to harm to entire groups, which may even contravene fundamental values. In such cases, we speak of discriminatory bias[28]. This is the case when AI systems lead to decisions or predictions that discriminate on the basis of race or ethnicity[29], political opinions, religion or beliefs, trade union membership, genetic condition, health status or sexual orientation.[30]

These errors may be related to the model (the algorithm chosen, the parameters used, the weights attributed to each variable, etc…); or to the data-

[25] Pasquale, F., *The Black Box Society: The Secret Algorithms That Control Money and Information*, Cambridge-London, Harvard University Press, 2015.

[26] LI, M., "Another Self-Driving Car Accident, Another AI Development Lesson", *Towards Data Science*, 20 November 2019, available at https://towardsdatascience.com/another-self-driving-car-accident-another-ai-development-lesson-b2ce3dbb4444.

[27] It may be that the use of certain AI algorithms to predict recidivism may lead to racial or gender bias, and predict a different probability of recidivism for males and females or for nationals and foreigners. See Tolan S., Miron M., Gomez E. and Castillo C., "Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia", *ICAIL 19: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*, 2019, pp. 83-92, available at https://doi.org/10.1145/3322640.3326705.

[28] EUROPEAN COMMISSION, 'White Paper on Artificial Intelligence - a European approach to excellence and trust', Doc. COM(2020) 65 final, pp. 13-15; FRA - EUROPEAN UNION AGENCY FOR FUNDAMENTAL RIGHTS, *Data quality and Artificial Intelligence- mitigating bias and error to protect fundamental rights,* Luxembourg, Publications Office*, 2019, https://fra.europa.eu/en/publication/2019/artificial-intelligence-data-quality.*

[29] Chivers, T., "Facial recognition… coming to a supermarket near you", 4 August, 2019, *The Guardian*, available at https://www.theguardian.com/technology/2019/aug/04/facial-recognition-supermarket-facewatch-ai-artificial-intelligence-civil-liberties.

[30] In general, see O'Neill, C., *Weapons of Math Destruction*, New York, Ramdon House, 2016.

sets with which it has been trained: the lack of volume, variety or quality of the data with which the system has been trained, or the existence of defects in the data pre-processing (duplications, generalisations, etc…).[31]

In the *Ethical Guidelines for Trustworthy AI*[32], the European Commission's High Level Expert Group on Artificial Intelligence rightly points out that humans and communities can only have confidence in technological developments and their applications if we have a clear and detailed framework to ensure their trustworthiness. Trustworthy Artificial Intelligence is based on three components: that it is ethical; that it is lawful; and that it is robust. In relation to the first of these components, the *Guidelines* identify four ethical principles that must be met to ensure that AI systems are developed, deployed and used in a trustworthy manner: respect for human autonomy, prevention of harm, fairness and explainability.[33]

In the opinion of the Expert Group:

> *"Explainability is crucial to gaining users' trust in AI systems and to maintaining that trust. This means that processes need to be transparent, that the capabilities and purpose of AI systems need to be openly communicated, and that decisions need to be explained - as far as possible - to parties who are directly or indirectly affected by them. Without this information, it is not possible to properly challenge a decision. It is not always possible to explain why a model has generated a particular outcome or decision (or what combination of factors contributed to it). Such cases, which are referred to as "black box" algorithms, require special attention. In such circumstances, other measures related to explainability (e.g., traceability, auditability and transparent communication about the system's performance) may be necessary, as long as the system as a whole respects fundamental rights. The degree of need for explainability depends to a large extent on the context and the seriousness of the consequences of an erroneous or inappropriate outcome.*

The principle of explainability in *the Guidelines* inspires a number of requirements introduced in the AIA for high-risk AI systems, which need to

[31] In this regard, it is worth recalling the controversy generated by a tweet by Yan Lecun, Facebook's chief researcher, regarding an AI model that had been used to transform Barack Obama into a white man, which ultimately led the researcher to leave the social network: *"LM systems are biased when data is biased. This face upsampling system makes everyone look white because the network was pretrained on FlickFaceHQ, which mainly contains white people pics. Train the *exact* same system on a dataset from Senegal, and everyone will look African"*. See "Yann LeCun Quits Twitter Amid Acrimonious Exchanges on AI Bias", *Synced*, 30 June 2020, available at https://syncedreview.com/2020/06/30/yann-lecun-quits-twitter-amid-acrimonious-exchanges-on-ai-bias/.

[32] INDEPENDENT GROUP OF HIGH-LEVEL EXPERTS ON ARTIFICIAL INTELLIGENCE (2018), *Guidelines for Trusted AI*, available at https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[33] *Idem*, p. 14.

be briefly referred to in order to put into context the right to an explanation. Thus, it is worth recalling that Article 10 imposes an obligation to implement appropriate data governance and management practices; Article 11 requires the existence of technical documentation; Article 12 requires the retention of records throughout the lifecycle of the AI system to ensure an adequate level of traceability of its operation; and Article 13, which requires AI systems to be designed in a way that ensures that they operate with a sufficient level of transparency for those responsible for deployment to correctly interpret and use their output information; finally, Article 14, which indicates that the AI system must be designed in such a way that the natural person entrusted with its oversight can adequately carry out his or her work.

The right to an explanation serves as a corollary to these requirements in that if a deployer is not able to provide the explanations requested by an affected person, this may be because the AI system does not comply with certain requirements, and therefore there is a breach of the Regulation that affects not only the individual who exercised the right, but society at large. This should be a sufficient indication for the supervisory authority to initiate investigation proceedings against the person responsible for the deployment and, where appropriate, against the provider.

3. Conditions for the exercise of the right to an explanation

The exercise of the right to an explanation is subject to certain conditions which, as will be seen, reduce its positive impact.

Firstly, the right to an explanation is only available in relation to decisions taken by those responsible for the deployment of high-risk AI systems listed in Annex III, with the exception of those listed in paragraph 2 above.

Thus, the AI systems in point 2 are those related to critical infrastructures. The importance of these infrastructures for the public authorities means that their management is entirely in the hands of the State.

AI systems that constitute safety components of products covered by the legislative acts listed in Annex I are also excluded. This is despite the fact that Article 86 states as one of the grounds for requesting explanations that the decision taken has a detrimental effect on "his health" or "his safety". It is conceivable that, as with the right to lodge a complaint in Article 85, the exclusion of these AI systems is based on the idea of disrupting as little as possible the functioning of the market surveillance authorities for the products covered by these legislative acts. In short, this means that the power to require a company to prove that the AI system incorporated in one of its products is "explainable" lies exclusively within the market surveillance authority. In any event, the person concerned may lodge a complaint with that authority in accordance with Article 85, but the authority is only obliged to take it into account.

Nor is this right enjoyed in relation to AI systems that are not considered high-risk, a circumstance that is justified by the same reasons that inform the regulation of these systems: their low impact on fundamental rights. And AI models of general use can also be considered excluded.

Finally, in the absence of an express exclusion, it can be said that the right is enjoyed in relation to AI models in general use. In such cases, in the absence of an explanation from the deployer, the complaint must be lodged with the AI Office.

Secondly, the right to an explanation provided for in Article 86 AIA is also not available in cases where:

(a) there are exceptions or restrictions to the obligation to provide explanations under Article 86.1 arising from Union or national law in compliance with Union law (paragraph 2 of the provision).

(b) the right to an explanation is otherwise provided for in Union law (paragraph 3). As will be explained in the following section, this exclusion mainly concerns the right to challenge an automated decision provided for in the GDPR.

Thirdly, the beneficiaries of this right should be specified. The provision refers to "any affected person subject to a decision which is taken by the deployer". Insofar as the wording does not differentiate, it must be understood that both natural and legal persons may exercise the right.

However, with regard to the former, it should be recalled that Article 86.3 indicates that the provision only applies "only to the extent that the right referred to in paragraph 1 is not otherwise provided for under Union law". It seems logical to state that a decision taken individually by an AI system affecting a natural person will be based on data relating to that person (personal data). This being the case, the individual should exercise the right to challenge an automated individual decision under Article 22 GDPR and not this right. As we will see in the following section, this circumstance may be relevant because the doctrine has stated that, at the very least, it is doubtful that the GDPR establishes a right to an explanation.

As regards the possibility for the person concerned to be represented by a collective protection body, there seems to be no room for doubt in this respect. More importantly with the amendment introduced by Article 110 AIA in Directive 2020/1828 which empowers associations to bring representative actions for the protection of the collective interests of consumers in the cases of non compliance with the AI Act.

Similarly, as we will see in section IV, according to Article 87, these persons can benefit from the protection offered by Directive 2019/1937 on claimant protection. This referral may have great relevance in those cases in which the person exercising the right is a worker or collaborator of the person responsible for the deployment who has first-hand knowledge of the errors that the AI system may make. However, in order to be able to exercise the right, the whistleblower must meet the conditions explained below.

Fourthly, the person concerned only enjoys a right to an explanation in relation to a decision that "produces legal effects or substantially affects him in the same way, so that he considers that it has a detrimental effect on his health, safety or fundamental rights".

To begin with, it should be noted that the CJEU[34] holds that the term "decision" must be interpreted broadly[35], which leads it to include the mere preparatory acts that serve to take the decision[36], which may have been carried out by different persons.

For the purpose of interpreting this condition, it is appropriate to consider the *Guidelines on automated individual decisions*. They indicate that a decision produces "legal effects" if it affects a person's legal rights, e.g., the freedom to associate with others, to vote in an election or to take legal action. A decision affecting a person's legal status (denial of a benefit granted by law, denial of admission to a country or denial of citizenship) or rights under a contract (cancellation of a contract) also produces a legal effect.

For its part, a decision that 'significantly affects a person in the same way' is to be understood as the same as a decision that 'significantly affects in a similar way' within the meaning of Article 22 GDPR. According to the *Guidelines*, these are decisions which, although they do not result in any change to the individual's legal rights or obligations, may affect him sufficiently to require protection. However, in such cases, the decision must "significantly" or "considerably" affect the individual. According to the *Guidelines*, this must be determined on a case-by-case basis, but in any event, the effects of the decision must be significant enough to be worthy of protection. This will be the case, for example, for decisions that affect a person's access to university studies; or deny them a job opportunity or place them at a disadvantage; or, where the AI system is used for marketing purposes, if the profiling of a person results in them being offered products or services at prohibitively high prices, which in practice prevents them from accessing them.[37]

[34] CJEU of 7 December 2023, C-634/21, *SCHUFA*.

[35] Paragraphs 44 y 45.

[36] Paragraphs 61 y 62.

[37] ART. 29 WORKING GROUP, *Guidelines…*, *op. cit.*

It should not be forgotten that the decision must have, on the person concerned, a "detrimental effect on his or her health, safety, or fundamental rights". The reference must be placed in line with the general objectives of the AIA, provided for in Article 1.1, but it does not seem that this condition can be an obstacle to the exercise of the right as it is easy to imagine that any decision taken on the basis of information provided by an AI system will affect one of these three objectives.

Fifth, in clear contrast to Article 22 GDPR, the right to an explanation in Article 86 AIA can be exercised in relation to a "decision that the controller takes on the basis of the results of an AI system". Its scope of application is different from that covered by the GDPR provision, which refers exclusively to 'decisions based solely on automated processing'.

This second scenario covers, for example, a decision to refuse a social benefit taken automatically by an AI system. On the other hand, Article 86 AIA concerns the decision of an official who, on the basis of information or a suggestion provided by an AI system, decides to refuse assistance.

It is difficult to believe that the intention of the institutions was to exclude automated individual decisions from the scope of Article 86. Therefore, despite the flawed wording, the right to an explanation can be interpreted as being exercisable in relation to both types of decisions.

Sixth, it is necessary to specify the requirements to be met by the explanation to which the deployer is obliged by this provision. These requirements relate to form ("clear and meaningful explanations") and content ("about the role the AI system played in the decision-making process and the main elements of the decision taken").

While the wording set out in Article 15.1(h) GDPR in relation to automated individual decisions is not identical ("*meaningful* information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject "), we consider it appropriate to take the *Guidelines on automated individual decisions* as a guide for interpreting these requirements.

In relation to the form, this document indicates that the person concerned should be informed in a simple manner and in a sufficiently comprehensive way so that he understands the reasons for the decision.

In relation to the content, meaningful information on the logic applied should be provided, not necessarily a complex explanation of the algorithms used or the disclosure of the entire algorithm[38]. In this respect it should be recalled that the requirements initially proposed by the Parliament have been

[38] ART. 29 WORKING GROUP, *Guidelines…*, *op. cit.*

lowered. Moreover, account should be taken of Recital 171 of the AIA which adds that the explanation should be able to "provide a basis for the persons concerned to *exercise their rights*".

As we will see in section 4, the determination of the content of "clear and meaningful explanations" also has implications from the point of view of intellectual property and the protection of confidential information of the provider and the deployer of AI systems.

Finally, reference should be made to the gaps and doubts raised by the regulation of this right. First, the provision does not set a time limit for the person responsible for the deployment to provide the requested explanations. This is important, for the purposes of the subsequent exercise of rights referred to in the aforementioned Recital 171. How much time should the person concerned allow to pass without receiving an explanation before taking legal action? Unfortunately, the European legislator has not taken as an example Article 12 GDPR, which obliges the controller to provide the requested information without delay and within a maximum period of one month, extendable for duly justified reasons.

Second, the provision does not establish what legal action can be taken. As explained above, Article 85 does not contain a genuine right to lodge a complaint with the market surveillance authority. If such a complaint is made, the complaint filed for failure to provide explanations or unsatisfactory explanations will be an additional element that the authority will take into account in determining whether or not to initiate an investigation against the deployer. Alternatively, the affected person may bring an action before the civil jurisdiction (in the case of a private entity) or the administrative jurisdiction (in the case of a public body). As indicated in the analysis of Article 85, the high cost and time-consuming nature of legal proceedings means that, in practice, this is only a viable option for entities representing the collective interests of consumers.

4. The relationship between the right to an explanation of Article 86 and the GDPR

As explained in the previous section, according to Article 86.3, the right to an explanation is applicable "only to the extent that the right […] is not otherwise provided for under Union law".

It is necessary to analyse whether this exclusion is applicable to the GDPR, as it is not clear whether a "right to an explanation" is contained therein or not. If it does, the GDPR would apply preferentially where the data subject(s) requesting an explanation are natural persons. In such a case, the right to an explanation under Article 86 AIA would lose some of its usefulness as only legal persons would benefit from it. On the other hand, if it were to be concluded that the GDPR does not provide for a right to an explanation, the usefulness of Article 86 AIA would be much greater, not only because it would also benefit natural persons, but also because it would reinforce the rights that natural persons have under Article 22 GDPR in relation to automated decisions.

The question of the existence of a right to an explanation in the GDPR has been widely discussed in the doctrine[39]. It is rightly argued that it is a multi-faceted concept. On the one hand, it refers to the explanations that the data subject is entitled to receive about the functioning of the system (i.e., the logic, meaning, and consequences flowing from it), or about the justification, reasons or individual circumstances that led to the adoption of a given decision. On the other hand, the right can be exercised before the automated decision has been taken (*ex ante*); or afterwards (*ex post*).[40]

The right to an *ex ante* explanation is adequately regulated in the GDPR. According to Article 5.1 of the GDPR, the controller (data subject) has the obligation to process data lawfully, fairly, and transparently. When such data are used for automated decision-making (including profiling), this implies an obligation on the controller to provide 'meaningful information about the logic applied and the significance and expected consequences of such processing for the data subject' (Article 15.1(h)).

But what happens when, despite complying with these obligations, a computer system makes an allegedly erroneous automated decision? Does the data subject have a right to an *ex post* explanation? For some, the above-mentioned information obligations and Article 22.3 GDPR introduce this right[41]. In particular, the latter provision gives the data subject a right "to express his or her point of view and to contest the decision ". However, none of these provisions expressly mention the right to an explanation, the reference to

[39] Vilasau I Solana, M. (2020), "La realización de perfiles y la salvaguardia de los derechos y libertades del afectado", in A. Cerrillo i Martinez and M. Peguera Poch, *Retos jurídicos de la inteligencia artificial*, Madrid, Aranzadi, 2020, pp. pp. 181 ff.

[40] Wachter, S., Mittelstadt, B., and Floridi, L., "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation", *International Data Privacy Law*, Vol. 7, No. 2, 2017, available at https://ssrn.com/abstract=2903469.

[41] Among others, Goodman, B. and Flaxman, S., "EU Regulations on Algorithmic Decision-Making and a "Right to Explanation", *AI Magazine*, vol 38, num. 3, 2017, available at 10.1609/aimag.v38i3.2741; Malgieri, G. / Comandé, G., "Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation", *International Data Privacy Law*, vol. 7, Issue 3, 2017, available at https://ssrn.com/abstract=3088976.

which can only be found in Recital 71, which speaks of "the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision".

The lack of legal endorsement and the non-binding nature of the recitals has led some authors to interpret that there is no right to an explanation in the GDPR[42]. Therefore, returning to the AIA, it could be understood that the exclusion in Article 86.3 will not be applicable to the GDPR, so that natural persons could benefit from the right to an explanation provided for in paragraph 1, in the terms analysed above. If this interpretation is upheld, the protection of individuals against automated individual decisions would be strengthened as the AIA grants them a right that does not seem to be covered by the GDPR.

However, the wording of Article 86.3 leads to another conclusion. This is that the provision does not exclude the application of the first paragraph when the right is provided for in another instrument of Union law, but when the right *is 'otherwise' provided for* under Union law. In my view, the right to an explanation can be interpreted as being 'otherwise' provided for in the GDPR when the automated decision relates to personal data. If this is the interpretation finally adopted, the practical usefulness of Article 86.1 is extremely limited in that only legal persons could be beneficiaries. The preparatory work also suggests that the legislator's intention was to exclude this right in those cases in which the GDPR is applicable. This explains why the version of the article introduced by the Parliament does not refer to EU law, but to "*Articles 13, 14, 15 and 22 of Regulation (EU) 2016/679".*

5. Limits to the right to an explanation: Intellectual property rights and confidential information

As is well known, AI systems are protectable by different categories of intellectual property. In particular, the models constitute computer programs that are copyrightable or patentable[43]; the weights and parameters used to train the system are copyrightable as databases or trade secrets[44]; and the dif-

[42] Wachter, S., Mittelstadt, B., and FLORIDI, L., *op. cit.*, p. 6.

[43] Muñoz Ferrandis, C. / Duque Lizarralde, M., "Open Sourcing AI: Intellectual Property at the Service of Platform Leadership" (January 26, 2022). Available at SSRN: https://ssrn.com/abstract=4018413

[44] Sousa E Silva, N, "Are AI models weights weights protected databases?", 18 January 2024, *Kluwer Copyright Blog*, available at https://copyrightblog.kluweriplaw.com/2024/01/18/are-ai-models-weights-protected-databases/.

ferent data sets that may be used to train the model may be protected, in themselves, by copyright or related rights, or as a whole as databases or, in the worst case and provided their confidentiality is ensured, as trade secrets[45]. The same applies to the results, which may be eligible for protection by copyright or related rights or as a trade secret. The beneficiaries of this protection may be the provider of the AI system, the deployer, or third parties.

It should be remembered that the exclusivity that these entities obtain over this software, the weights and training parameters, and over the data gives them a competitive advantage in the market that is worthy of protection by the legal system.

In this circumstance, when the right to an explanation is exercised, a conflict of interests arises: those of those responsible for the deployment to provide as little information as possible in order to preserve the intellectual property rights and the confidentiality of the AI system's data; and those of the person concerned to obtain as detailed an explanation as possible of the decision that the AI system has taken about him or her. In my view, it is not only in the interest of the data subject but also in the interest of society as a whole to obtain as detailed an explanation as possible, as this helps to detect errors in AI systems which, after all, are of general benefit. Given this conflict, the question arises: is the deployer obliged to disclose proprietary or trade secret information where this is necessary to provide a clear and meaningful explanation of the role that the AI system has played in the automated decision-making process?

Neither Article 86 AIA nor the related recitals include any precision as to how to answer this question. However, an analysis of other provisions of the Regulation that address the issue of the treatment of intellectual property rights and confidential information leads to a negative answer.

Thus, in all provisions where the provider or other participant in the value chain is obliged to provide information on the AI system, it is stated that this obligation is: "without prejudicee to the need to observe and protect intellectual property rights, confidential business information and trade secrets in accordance with Union and national law" (Article 25.5 on the obligations of economic operators involved in the AI value chain, including deployers; Article 53.1(b) on the obligations of providers of general purpose AI models).

The same obligation falls on the Commission to publish the list of general-purpose AI models with systemic risk referred to in Article 52.6.

Finally, in general, Article 78 obliges 'market surveillance authorities, noti-

[45] Extensively, in Lopez-Tarruella Martinez, A, *Propiedad intelectual e innovación basada en los datos*, Madrid, Dykinson, 2021.

fied bodies and any other natural or legal person involved in the enforcement of this Regulation shall, in accordance with Union and national law, respect the confidentiality of information and data obtained in the exercise of their tasks and activities in such a way as to protect, in particular (a) intellectual and industrial property rights and confidential business information or business secrets of a natural or legal person'. Furthermore, in relation to the authorities, paragraph 2 states that they 'shall only request data which are strictly necessary for the assessment of the risk presented by AI systems and for the exercise of their powers in compliance with this Regulation and Regulation 2019/1020'.

Therefore, although Article 86 does not expressly provide for it, from the set of provisions analysed it appears that the deployer is not obliged to provide information that may be considered a trade secret or protected by intellectual and industrial property rights, when providing explanations on the decision taken on the basis of the results provided by the AI system. And, in the event that the deployer deems it necessary to provide it, Article 78.1 obliges the person concerned to respect the confidentiality of the information and data obtained.

However, it would have been appropriate to introduce a precision similar to that contained in Recital 63 GDPR: the need to preserve their intellectual property rights and the secrecy of confidential information cannot result in the rejection of the request for explanations. In other words, the request must be complied with but only by providing information that does not harm the interests of the controller or third parties.

IV. The Role of Complainants and Collective Interest Associations

As indicated in the introduction, the regulation of the right to lodge a complaint with a market surveillance authority in Article 85, and of the right to an explanation of decisions taken individually in Article 86, is complemented by Articles 87 and 110. The former contains a cross-reference to Directive 2019/1937 on the protection of complainants. The second amends the Annex to Directive 2020/1828 to ensure that associations representing the collective interests of consumers can bring actions for breach of the Regulation. Both provisions therefore relate to standing to exercise these rights. For reasons of exposition, it has been preferred to treat the two provisions separately in this section.

In view of the benefits for the public interest, Directive 2019/1937 grants protection to persons who report breaches of EU law. Such complaints enables effective detection, investigation and prosecution of such infringements, thereby improving transparency and accountability[46].

Although the AIA can be considered as one of the instruments falling within the scope of the Directive by the reference in Article 1.1(c) to "internal market infringements", the European legislator has preferred to remove any doubt in this respect with an express reference in Article 87 AIA:

Directive (EU) 2019/1937 shall apply to the reporting of infringements of this Regulation and the protection of persons reporting such infringements.

In particular, the Directive obliges Member States to take the necessary measures to ensure that the identity of the complainant is not disclosed without his or her express consent (Article 16), and to prohibit all forms of retaliation against such persons (Articles 19 and 21). Support measures for the complainant, such as the provision of comprehensive and independent information and advice, effective assistance by the authorities against retaliation and legal aid, must also be put in place (Article 20). Support measures should also be put in place for other persons affected by their relationship with the complainants (Article 22).

Such whistleblower protection can go a long way in helping to promote compliance with the AIA. In this regard, it is worth recalling that in the past, the public and the authorities have learned of blatant breaches of applicable regulations by technology companies thanks to disclosures made by employees or collaborators of these entities. This was the case, for example, with the *Cambridge Analytica* scandal.

Similarly, recital 8 of the Directive reminds us, in relation to the safety of products placed on the internal market (as may be the case for products incorporating AI systems), that "undertakings operating in the manufacturing and distribution chains are the main source of evidence, so that information from whistleblowers in these undertakings has a high added value as they are much closer to information on possible abusive and illegal manufacturing, import or distribution practices relating to unsafe products. Consequently, there is a need for whistleblower protection to be introduced in relation to the safety requirements applicable to products regulated by Union harmonisation legislation, as set out in Annexes I and II of Regulation (EU) 2019/1020."

This recital also indicates that market surveillance authorities or judicial authorities will be obliged to guarantee the complainant of an infringement of the AIA the protection afforded to him by the national rules implementing this Directive.

In addition, the inclusion of AIA in the Annex to Directive 2020/1828

[46] Recital 1.

empowers bodies for the protection of the collective interests of consumers (so-called "qualified entities") to bring representative actions against acts of traders in breach of the provisions of AIA. This would strengthen consumer confidence and empower consumers to exercise their rights, contribute to fairer competition and create a level playing field for businesses operating in the internal market. In the AIA environment, the existence of these mechanisms should incentivise providers and other operators in the AI value chain to properly comply with the requirements and obligations of the Regulation.

According to the Directive, Member States must ensure that at least an effective and efficient procedural mechanism for representative actions for injunctions and redress is available to consumers at Union and national level. The reference to "consumers" implies that Member States are not obliged to ensure such mechanisms when infringements of EU law harm natural or legal persons who are considered to be entrepreneurs.

In the case of the AIA, unlike in the rest of European digital laws, this effective procedural mechanism necessarily involves the exercise of legal action since, as explained in the first section, there is no real right to file a complaint with the market surveillance authorities. It is true that, in these cases, Article 9 Regulation 2019/1020 establishes the power of market surveillance authorities to agree with "organisations representing economic operators or end users" to carry out joint activities with a view to encouraging compliance or detecting cases of non-compliance. In the same vein, in relation to legal actions, the Directive ensures that entities qualified in one Member State can bring representative actions in another Member State, and should be able to join forces to bring a single action before a single forum.

It should be noted how effective the implementation of the Directive can be in promoting effective compliance with the AIA. Suffice it to recall the effectiveness of the actions brought by the entities *NOYB* or *La Quadrature du Net* to promote compliance with the GDPR; and the actions brought recently in the Netherlands, under the WAMCA[47], against technology giants such as Apple, Google, or Tik Tok requesting the adoption of injunctions for breach of the Digital Services Regulation, and of European *antitrust* law.[48]

[47] Class Action Settlement Act (*Wet collectieve afwikkeling massaschade*).

[48] X. Kramer, "International tech litigation reaches the next level: collective actions against TikTok and Google", Conflict of laws, 12 March 2024, available at https://conflictoflaws.net/2024/international-tech-litigation-reaches-the-next-level-collective-actions-against-tiktok-and-google/.

V. Conclusions

Section 4 ("Remedies") of Chapter IX ("Post-market surveillance, exchange of information, and market surveillance") of the AIA was introduced by the European Parliament at the legislative stage in order to guarantee individuals certain rights when they are affected by the use of AI systems. The right to lodge a complaint with a market surveillance authority and the right to an explanation of decisions taken on an individual basis are therefore to be welcomed. However, its final wording has substantially emptied the content initially attributed to it by the Parliament.

Thus, on the one hand, the former does not regulate a right to lodge a complaint in the same sense as the GDPR or other European digital laws. Rather, what it regulates is a right to file a petition with the market surveillance authority, which will take this into account when determining whether to initiate investigative operations against the provider or deployer of the AI system. Furthermore, the submission of such petitions is hampered by the complex distribution of competences between the market surveillance authorities resulting from Article 74. While this provision allows for a certain degree of concentration of competence to deal with such complaints within AESIA, this solution is not without its problems.

On the other hand, in relation to the second, it should be noted that the right is only enjoyed in relation to certain categories of high-risk AI systems: those listed in Annex III (with the exception of point 2). Moreover, if our interpretation of the relationship of Article 86.3 AIA with the GDPR is correct, this right can only be exercised by legal persons. In the case of natural persons, insofar as the automated decision making that affects them should necessarily have been carried out on the basis of data allowing them to be identified, the obtaining of explanations must be exercised through the channels provided for by the GDPR.

In short, we are faced with a well-intentioned regulation of remedies in Articles 85 to 87, which, due to last-minute negotiations in the trialogue, has lost much of its useful effect to the detriment of the interests of individuals and the objectives set out in Article 1 of the AI Act.

ACCESS TO DOCUMENTS AND CONFIDENTIALITY IN THE ARTIFICIAL INTELLIGENCE ACT

Gabriele Vestri

PhD in Law, Founder and President of the Public Sector and Artificial Intelligence Observatory.

I. Introduction

As the title of this contribution suggests, our purpose is to analyse two of the essential pillars of the AIA. This regulation represents the latest advance made by the powers of the European Union, under the leadership of the Spanish representation last December 2023. A regulation that finally responds or attempts to respond to what Salazar García has called "technological shock", which in some way unites technological advances with the fear that they produce[1]. Precisely in order to respond to this change, the AIA creates a scaffolding of rules, sometimes complex, which, from a horizontal perspective, is not limited to specific sectors but aims to mitigate the harmful effects of Artificial Intelligence.[2]

Thus, in this contribution we refer specifically to the provisions of the AIA related to access to documentation and confidentiality. In this context, our aim is to unravel the legal, administrative and practical complexities surrounding certain critical issues in the regulatory development of Artificial Intelligence in the EU. We will focus in particular on Articles 77 and 78 of the AIA, although it should be noted that these rules raise certain ramifications that we will attempt to address and analyse.

In this scenario, it is necessary to introduce some concepts that will help us to understand and analyse the scope of the issues addressed, which, as can be imagined, are not without complications. Furthermore, it should be noted that the analysis we propose assumes the definitions present in the AIA, referring to them without being repeated in this context.

It is useful, however, to make a universal approximation to certain notions that, at the very least, allow us to trace the path to follow in our contribution.

In addressing access to documentation and confidentiality, we are dealing

[1] Salazar García, I. "Privacidad e inteligencia artificial: ¿es posible su convivencia?" in Arellano Toledo, W. (Director), in "*Derecho, Ética e Inteligencia Artificial",* Tirant lo Blanch, (2023), p. 181.

[2] In this sense, see: Barrio Andrés, M. "Inteligencia artificial, Internet de las cosas y *blockchain*" in Montero Pascual, J.J. (Coordinator), "*Digitalización y derecho. Curso de Derecho digital*", Tirant lo Blanch, (2024), p. 266.

with the criteria of transparency. AIA is based on the indispensable principle of transparency. To ensure that users, authorities, and also citizens fully understand the impact and functioning of Artificial Intelligence systems, there is an obligation to provide detailed documentation - transparency which, in a more general approach, is addressed in Article 13 of the AIA, which we recommend reading. Well, this documentation must not only be clear and understandable, but also accessible, reflecting a commitment to the informed participation of all interested parties. Precisely within the scope of the EU Regulation and as Cotino Hueso rightly points out, we must take into account the typology of information so that "the user or consumer of the system (and the technicians who implement it), can manage the AI system correctly, fulfil their obligations and supervise them. The different nature of users, importers and distributors must be taken into account"[3]. This is not a trivial issue, and it is our duty to point out that, although there are some differences in approach, some Member States have already put forward arguments in which access to information has played a central role. It is enough to remember the Spanish case known as Bono social-Fundación Civio, which focused precisely on the denial of access to information from the Bosco system[4]. In the same vein, several Italian court rulings have highlighted the need for access to information and the inherent comprehensibility.[5]

This approach leads to what is known in national legislation as the right of access. A prominent aspect is the recognition of the right of users to access relevant documentation. This not only strengthens the position of users in an increasingly AI-driven world, but also promotes the accountability of the providers of these systems. Transparency - through access - becomes a means to empower and ensure informed autonomy.[6]

Naturally, such access must be subject to certain limitations. In the interests of a balanced approach, the AIA also sets limitations on full disclosure

[3] Cotino Hueso, L. "Transparency and explainability of Artificial Intelligence and "company" (communication, interpretability, intelligibility, auditability, testability, testability, testability, simulability…). For what, for whom and how much" in Cotino Hueso, L. Claramunt Castellanos, J. (Coordinators). "*Transparencia y explicabilidad de la inteligencia artificial*". Tirant lo Blanch, (2022), p. 46.

[4] See Vestri, G. "El acceso a la información algorítmica a partir del caso bono social vs. Fundación ciudadana Civio" in *Revista General de Derecho Administrativo*, n.º 61, (2022), pp. 1-24.

[5] For an overview of the orientation of Italian jurisprudence see: Vestri, G. "Sistemi algoritmici e principio di buona amministrazione algoritmica" in *Rivista Diritto di internet*, n.º 2, (2023), pp. 373-382.

[6] On algorithmic transparency See: Vestri, G. "La inteligencia artificial ante al desafío de la transparencia algorítmica. An approach from a legal-administrative perspective". *Revista Aragonesa de Administración pública*, n.º 56, (2021), pp. 368-398.

of information, recognising that certain details may compromise public security, privacy or intellectual property rights. The exceptions and limitations seek to safeguard other fundamental values without completely stripping the regulation of its transparent nature.

The AIA also deals with confidentiality, which, in our view, must also be considered in relation to data protection. Indeed, in the area of confidentiality, the AIA takes into account the sensitive issue of data protection. Developers of Artificial Intelligence systems handle strategic, commercial, and research information, which makes it imperative to guarantee its confidentiality so as not to compromise competitiveness and innovation. In the same context, it is of utmost importance to proceed with the corresponding conformity assessment of the different forms of confidentiality. Conformity assessments, the cornerstone of the regulatory framework, are also subject to the prism of confidentiality. This approach seeks to preserve intellectual property and trade secrets associated with the assessment processes, while ensuring the integrity of the regulatory system.

Of crucial importance is also the approach to national security and strategic limitations. In this regard, and in recognition of the need to protect national security, the AIA incorporates provisions that allow for limitations on the disclosure of information that could endanger the security of the Member State and the Union. This nuance highlights the EU's awareness of the need to balance technological innovation with national security.

To conclude this introductory part, and as is probably already understood, the EU AIA is presented as an ambitious regulatory framework, underpinned by sound principles of access (and thus transparency) and confidentiality. This cautious and balanced approach reflects the EU's commitment to ethics and responsibility in the development of Artificial Intelligence. As we move forward in studying the effects of Artificial Intelligence, it is imperative to further scrutinise these provisions, assessing their implementation and adapting them to a constantly evolving technological environment. The convergence of technology and law calls for continuous vigilance and critical reflection, and in this sense, we find ourselves at the epicentre of a fascinating and challenging legal terrain. This analysis will always be carried out from a critical approach, as we will try to break down in this contribution and in the knowledge that the European standard, apart from its strategic importance, could have been even more ambitious than it really is. The trend in AIA is to develop an ecosystem of excellence and also to create an ecosystem of trust[7]. Perhaps only time will allow us to assess the impact of AIA.

[7] See: Muñoz García, C. "*Regulación de la inteligencia artificial en Europa. Incidencia en los*

II. Analysis of the content of Article 77 of the AIA

It is important to note that Article 77 of the AIA falls within the framework of what the same regulation establishes as: "Powers of authorities protecting fundamental rights". Now, although the title of the rule is undoubtedly eloquent, it is perhaps correct to point out that the legal provision in question is configured as a rule that grants specific powers to the national supervisory authority in relation to datasets used in activities linked to Artificial Intelligence or automated systems. Having said that, there is no doubt that access to documents, access to the Artificial Intelligence algorithm, is very closely related to fundamental rights, at least from the perspective that an Artificial Intelligence system poses individual and societal risks that may precisely endanger fundamental rights.[8]

According to the text, the national supervisory authority, in the exercise of its powers and upon submission of a duly substantiated request, has the right to obtain full access to the datasets used in the training, validation, and testing stages by the provider or, where applicable, the deployer. This granting of access is limited to those datasets that are relevant and strictly necessary for the purposes for which the access request was made. The implementation of such access must be carried out using appropriate technical means and tools that ultimately allow structured and proactive access.

It is imperative to emphasise that this regulatory provision aims to ensure transparency and effective oversight of activities associated with Artificial Intelligence, recognising the importance of data sets in the evaluation and control of automated systems. The reasoned submission of the request and the limitation to strict relevance and necessity of the data seek to balance the oversight authority with the protection of confidentiality and other legitimate rights of the providers or deployers. All this, and paraphrasing Cotino Hueso, is valuable information that is closely linked to the principle of proportionality, serving as an alternative both in general for the public authority and, in particular, when there are restrictions or impacts on fundamental rights.[9]

Likewise, and in a broader perspective, the Spanish Data Protection Agency has also pronounced, naturally within the scope of its subject matter,

regímenes jurídicos de protección de datos y de responsabilidad por productos" Tirant lo Blanch, (2023), p. 36.

[8] In this sense, see: Presno Linera, M.Á. "*Derechos fundamentales e inteligencia artificial*". Marcial Pons, (2022), pp. 23-24.

[9] See Cotino Hueso, L. "Qué concreta transparencia e información de algoritmos e inteligencia artificial es la debida" in *Revista Española de la Transparencia*, n.º 16 primer semestre enero-junio (2023), p. 30.

on the impact of the AIA on the issue addressed here. In this regard, it points out that: "when AI systems are included in, or are means of, a processing of personal data, controllers must obtain sufficient information about them to meet their various GDPR compliance obligations. These include transparency to enable the exercise of rights, to comply with the principle of active accountability, to meet the requirements of the GDPR Supervisory Authorities in relation to their investigatory powers, and the same for certification bodies and code of conduct monitoring"[10].

Subsequently, the rule sets out a detailed legal framework for the national supervisory authority in the context of high-risk Artificial Intelligence systems. It is highlighted in the rule that, in necessary situations and upon submission of a duly substantiated request, the national supervisory authority has the right to access the trained model and the training model of an Artificial Intelligence system, as well as the relevant parameters of these models. Such access is granted after all other reasonable means of verifying the compliance of the high-risk Artificial Intelligence system, including those referred to in the previous paragraph, have been exhausted and shown to be insufficient. The conformity assessment is intended to ensure that the Artificial Intelligence system complies with the pre-established requirements.

It is of utmost importance to note that all information obtained during this procedure and in accordance with Article 78 is considered confidential information. Such information is subject to the European Union's rules on the protection of intellectual property and trade secrets. It is also specified that this information will be deleted once the investigation for which it was requested has been concluded.

The introduction of paragraph 2a emphasises that the procedural rights of the operator under Article 18 of Regulation (EU) 2019/1020 are not affected by the above provisions. In other words, it ensures that the operator of the high-risk Artificial Intelligence system retains its procedural rights during the conformity assessment process carried out by the national supervisory authority.

The text further provides that national public authorities or bodies vested with responsibility for supervising compliance with obligations under EU law, in particular those relating to fundamental rights and non-discrimination in the use of high-risk Artificial Intelligence systems, have the power to request and obtain access to any documentation generated or stored under the

[10] Spanish Data Protection Agency, "*Artificial Intelligence: transparency*", at https://www.aepd.es/prensa-y-comunicacion/blog/inteligencia-artificial-transparencia [Accessed 28 December 2023].

regulation itself. Such access is granted whenever it is indispensable for the execution of their powers within the limits of their territorial competence.

It is imperative to underline that access to the documentation must be in an accessible and therefore understandable language and format. Furthermore, when making such a request, the relevant public authority or body is obliged to inform the market surveillance authority of the Member State concerned about the request. However, by extension, and as Belloso Martín points out, we should perhaps aim for the algorithm to be not only explainable but also fair, and this is the real challenge.[11]

In short, the text under review seeks to empower national authorities responsible for safeguarding fundamental rights to obtain the relevant documentation in the field of high-risk Artificial Intelligence systems, thus ensuring effective control and supervision of compliance with the obligations arising from EU law in this area.

The Article goes on to provide that within three months of the entry into force of the Regulation, each Member State shall identify the public authorities or bodies referred to in paragraph 3 of the relevant legislation[12]. The identification of these entities must be disclosed by publishing a list on the website of the national supervisory authority of the respective Member State. Member States are also obliged to notify this list both to the Commission and to all other Member States and to keep it up to date.

In simple terms, this paragraph sets a specific deadline for each Member State to identify and publish the public authorities or bodies referred to in the regulation. The disclosure of this information on the website of the national supervisory authority, together with the notification to the Commission and other Member States, aims to ensure transparency and effective communication between Member States and the Commission in the context of the implementation and application of the Regulation.

The text establishes a legal procedure when the available documentation, as set out in paragraph 3 of the relevant regulation, proves to be insufficient to determine whether a breach of obligations under Union law to protect fundamental rights has occurred in the context of high-risk Artificial Intelligence systems.

However, in situations where the documentation specified in paragraph

[11] Belloso Martín, N. "Sobre fairness y machine learning: el algoritmo ¿puede (y debe) ser justo?" in *Anales de la Cátedra Francisco Suárez* n.º 57, (2023), p. 3. DOI: https://doi.org/10.30827/acfs.v57i.25250

[12] It is known that Spain already has a Spanish Agency for the Supervision of Artificial Intelligence (AESIA).

3 does not provide sufficient information to verify whether there has been a breach of obligations under Union law aimed at protecting fundamental rights in the area of high-risk AI systems, the public authority or body referred to in the same paragraph has the power to submit a reasoned request to the market surveillance authority. The request aims at organising a verification of the high-risk AI system by technical means.

Finally, the market surveillance authority will, in turn, carry out the necessary tests with the close involvement of the requesting public authority or body. This process should be carried out within a reasonable time after receipt of the request. In essence, this mechanism allows the public authority or body, where existing documentation is insufficient, to request the market surveillance authority to carry out technical tests to assess the compliance of the high-risk AI system with the obligations under Union law related to fundamental rights.

Understandably, access to documentation seems to become what national legal systems usually define as "public information" so that this criterion should also be considered within Article 77.[13]

III. Analysis of the content of Article 78 of the AIA

The purpose of the aforementioned provision, i.e., Article 78, is to delineate the parameters relating to confidentiality, which are considered to be meticulously precise in their wording. For the sake of precision and as a general approximation, it should be noted that confidentiality refers to the protection and preservation of sensitive or confidential information handled, in this case, in digital environments. In this context, confidentiality stands as a fundamental pillar for safeguarding business data, trade secrets, personal information and other digital assets crucial to the parties involved. Indeed, systems to ensure confidentiality focus on establishing and enforcing legal and technical measures, such as confidentiality agreements, data encryption and restricted access policies, to ensure that confidential information is not disclosed or misused. Confidentiality, in this sense, not only protects the interests of the parties involved, but also contributes, or at least attempts to contribute, to building trust in the digital environment, promoting innovation, and technological development in a secure manner.

[13] On the subject in question we recommend: Gutiérrez David, M.E. "Administraciones inteligentes y acceso al código fuente y los algoritmos públicos. Conjuring risks of decisional black boxes" in *Derecom,* n.º 30. Nueva Época. March-September, (2021) pp. 159-160.

In this regard, the first paragraph lays down provisions concerning the confidentiality of information and data in the context of the implementation of the AIA. The text prescribes that the Commission, market surveillance authorities and notified bodies, as well as any natural or legal entity involved in the implementation of the AIA, are bound to observe confidentiality with regard to the information and data they obtain in the performance of their tasks. This safeguarding of confidentiality should be in accordance with Union or national law.

Special emphasis is given to safeguarding intellectual property rights, confidential business information, trade secrets, including source code, except in those cases covered by Directive 2016/943 on the safeguarding of undisclosed know-how and business information.

It also lists a number of purposes for which confidentiality protection is required:

a) The effective implementation of this Regulation, in particular for the purposes of inspections, investigations or audits.

b) Consideration of public interests and national security.

(c) The integrity of information classified in accordance with Union or national law.

In other words, an obligation is introduced for the authorities involved in the enforcement of the Regulation to require only the strictly necessary data to assess the risk posed by the AI system and to exercise their powers in line with the relevant Regulations. In addition, it underlines the need to implement appropriate and effective cybersecurity measures to protect the security and confidentiality of the information and data obtained. It imposes the obligation to delete the data collected once they are no longer necessary for the purpose for which they were requested, in accordance with the applicable national or European legislation, with express reference to Regulation 2019/1020.

The rule in question, in this case the second paragraph, imposes limitations on the disclosure of information which has been exchanged on a confidential basis between national competent authorities and between these authorities and the Commission, in the context of the use of high-risk Artificial Intelligence systems, specifically those indicated in points 1, 6 and 7 of Annex III.

Thus, without prejudice to the provisions of paragraphs 1 and 2, a statement is introduced to the effect that the above provisions shall not affect the provisions of paragraphs 1 and 2 of the relevant policy instrument.

Information confidentially exchanged between national competent authorities and the Commission: refers to confidential data shared between the national competent authorities of the Member States and the European Commission. Shall not be disclosed without prior consultation: This prohibits the disclosure of such information without prior consultation of the originating competent national authority and the user.

Where high-risk AI systems referred to in points 1, 6 and 7 of Annex III are used by law enforcement, border control, immigration or asylum authorities: The restriction applies specifically when these high-risk AI systems are used in contexts linked to law enforcement, border control, immigration or asylum authorities.

Also, where such disclosure could endanger the interests of public and national security: establishes the criterion that disclosure can only be avoided if it is considered that this action could endanger the interests of public and national security. This exchange of information shall not cover sensitive operational data in relation to the activities of law enforcement, border control, immigration or asylum authorities: delimits the information exchanged, excluding sensitive operational data related to the activities of the above mentioned authorities.

The fourth paragraph of the draft agreement introduces a regulatory safeguard, stating that paragraphs 1, 2 and 3 shall not affect certain specific rights and obligations of the Commission, Member States, their relevant authorities and notified bodies. The non-impact concerns in particular the exchange of information, the dissemination of alerts, cross-border cooperation and the obligations of the parties concerned in the context of the enforcement of Member States' criminal law.

Paragraphs 1, 2 and 3: This refers to Sections 1, 2 and 3 of the legislation in question, which contain specific provisions.

Nor shall they affect the obligations: states that these provisions shall not modify or affect the rights and obligations of the aforementioned subjects.

Of the Commission, Member States and their competent authorities, as well as notified bodies: details the subjects whose rights and obligations will not be affected, including the European Commission, Member States and their respective competent authorities, as well as notified bodies.

As regards the exchange of information and dissemination of alerts, including in the context of cross-border cooperation, the paragraph in question limits the non-impact to situations related to the exchange of information and dissemination of alerts, in particular in the context of cooperation between different jurisdictions.

It is also clarified that the obligations of the actors involved (stakeholders) to provide information in accordance with the criminal law of the Member States will also remain unaffected.

In other words, it operates as a non-action clause, ensuring that certain specific rights and obligations related to information exchange, dissemination of alerts, cross-border cooperation, as well as obligations under criminal law, are not altered by the provisions contained in the aforementioned paragraphs.

Finally, the last paragraph of Article 78 provides for the possibility of exchange of confidential information between the Commission and the Member States of the European Union, as well as the regulatory authorities of third countries. The implementation of such an exchange is conditional upon necessity and must be carried out in accordance with the specific provisions of international and trade agreements. Furthermore, it is stressed that this exchange can only take place with those third country regulatory authorities with which bilateral or multilateral confidentiality arrangements have been concluded which ensure an adequate level of protection of confidential information.

In strictly technical-legal terms, a certain scenario arises.

The Commission and the Member States may exchange: establishes the power of the Commission and the Member States to carry out the exchange of information.

Where necessary and in accordance with relevant provisions of international and trade agreements: makes the implementation of such exchange conditional on necessity and prescribes that it must be in accordance with the specific provisions of international and trade agreements.

Furthermore, it is underlined that such exchange can only take place with those third country regulatory authorities with which bilateral or multilateral confidentiality agreements have been concluded that ensure an adequate level of protection of confidential information: it emphasises that communication of information can only take place with third country regulatory authorities that have bilateral or multilateral confidentiality agreements in place, thus ensuring a sufficient level of protection for confidential information.

IV. Conclusions

The above describes two of the legal provisions concerning the regulation of Artificial Intelligence in the context of the European Union. The rules under study focus on the national supervisory authority and its specific powers to access datasets used in activities linked to Artificial Intelligence, with the purpose of ensuring transparency and effective supervision of activities associated with Artificial Intelligence, recognising the importance of datasets in the evaluation and control of automated systems.

First, it establishes the right of the supervisory authority to access relevant and strictly necessary data sets for the purposes of training, validation, and testing of Artificial Intelligence systems. This measure is presented as a balance between the supervisory authority and the protection of confidentiality and other legitimate rights of providers and deployers.

The standard subsequently focuses on high-risk Artificial Intelligence systems, giving the supervisory authority the right to access the trained model and the training model, as well as the relevant parameters. The conformity assessment aims to ensure that these systems comply with the requirements set out in the legal framework. The information obtained during this process is considered confidential and is subject to intellectual property and trade secret regulations, with an obligation to delete it once the investigation is concluded.

The guarantee of procedural rights of the operator during the conformity assessment is highlighted. In addition, national authorities are given the power to require access to documentation related to compliance with obligations under Union law in the field of high-risk Artificial Intelligence systems, ensuring effective control.

The text sets a deadline for Member States to identify and disclose the public authorities or bodies responsible for monitoring the obligations of Union law in this area. This disclosure is intended to ensure transparency and effective communication between the Member States and the Commission.

Article 77 sets out detailed parameters on confidentiality, highlighting the protection of intellectual property, confidential business information and trade secrets in the context of the implementation of the Regulation. It emphasises the need to only require strictly necessary data and the implementation of cybersecurity measures. The limitation to the disclosure of confidential information exchanged between national authorities and with the Commission is also established, specifically in the context of high-risk AI systems.

The Article addresses situations where the available documentation is insufficient by allowing the market surveillance authority to carry out technical tests in cooperation with the requesting public authority or body. This mechanism ensures a proper assessment when existing documentation is not sufficient.

In short, the text seeks to balance effective oversight of Artificial Intelligence with the protection of confidentiality and the rights of providers. It establishes a detailed legal framework for high-risk Artificial Intelligence systems, ensures transparency and communication between authorities and defines clear parameters for confidentiality and information sharing.

The provisions examined, as well as the entire text of the AIA, represent only a first step in the direction of regulating Artificial Intelligence. This means that, understandably, it is necessary to await the actions of the various Member States before proceeding to an analysis, in the form of an impact test, in order to determine whether European legislation, together with national legislation, has succeeded in establishing a proactive regulatory structure and environment in the field of Artificial Intelligence.

Also, at its core, and not so core, the AIA is very much like a kind of trade treaty. This should not necessarily be understood in a negative way, but we must understand how difficult it can be for its principles to be directly reflected in people's lives. Rather, the AIA establishes rules of commercial coexistence between professional actors. This is why we insist that the Member State's national implementation of the AIA will be crucial. It is at this point that people, citizens, will be able to see and feel how the rules regulate their relationship with Artificial Intelligence.

In a context where the evolution of Public Administrations is perpetually shifting from legal, regulatory, and jurisprudential perspectives, thereby influencing every organisation and activity, Public Administration must inevitably incorporate extensive and diverse knowledge, both specific and interdisciplinary, pertaining to organisational structure and the execution of their competencies.

This is necessary to achieve an administrative action that is increasingly efficient and impartial, as well as to understand the needs of society and social issues and the expectations of citizens towards public entities in the administered community.

In this perspective, the CERIDAP series, which was established in conjunction with the homonymic Interdisciplinary Research Centre on Public Administration Law at the University of Milan and is closely linked to the CERIDAP Journal (https://ceridap.eu), aims to provide comprehensive analyses of subjects that are relevant to all three pillars of administration (organisation, activities, and judicial protection) and are conducted from a multidisciplinary perspective.

The CERIDAP series indeed positions itself as a place for in-depth study and research on issues related to the functioning of Public Administration, from the perspective of so-called good administration.

ISBN 979-12-235-0281-5

9 791223 502815