CYBERSECURITY & ARTIFICIAL INTELLIGENCE: THE PLAYBOOK ON EUROPEAN GOVERNANCE, STRATEGIES, AND LAW

Edited by

Annita Larissa Sciacovelli
Carmelo Dimauro
Mattia Siciliano

Editoriale Scientifica

CYBERSECURITY & ARTIFICIAL INTELLIGENCE: THE PLAYBOOK ON EUROPEAN GOVERNANCE, STRATEGIES, AND LAW

Edited by

Annita Larissa Sciacovelli
University of Bari Aldo Moro

Carmelo Dimauro
LHC – Luxembourg House of Cybersecurity

Mattia Siciliano
Luiss Business School

Published under the patronage of
LHC – Luxembourg House of Cybersecurity

EDITORIALE SCIENTIFICA
Napoli











Open-access Playbook, funded by the European Commission under the Jean Monnet Module "Cybersecurity in the Digital Era: European Perspectives – CDeEP 2025-2027"
University of Bari Aldo Moro, Italy
ERASMUS-JMO-2024-HEI-TCH-RSCH | Proposal n. 101176128

This publication was produced with the financial support of the European Union under the Jean Monnet Programme. The contents of this document are the sole responsibility of the authors and can under no circumstances be regarded as reflecting the position of the European Union or the European Commission.

All chapters in this volume underwent blind peer review prior to publication.

© Copyright 2026 Editoriale Scientifica s.r.l.
Via S. Biagio dei Librai, 39
Palazzo Marigliano
80138 Napoli

Tutti i diritti (traduzione, adattamento) sono riservati per tutti i Paesi.
La riproduzione, anche parziale, e con qualsiasi mezzo
(compresi microfilms e fotostatiche) è vietata.

ISBN 979-12-235-0849-7

TABLE OF CONTENTS

Foreword 7

Preface of the editors 9

Annita Larissa Sciacovelli, The Normative and Institutional Architecture of EU Cybersecurity and Cyber Defence: Between Market Integration and National Sovereignty 13

Emmanuel Kessler, Cybercrime & LEA 25

Pascal Steichen, National Cybersecurity Strategies: Sovereignty, Governance Models, and Public-Private Partnerships 37

Raffaele Marchetti, Paolo Spagnoletti, The Geopolitical Compass of Cyberspace: US-PRC Rivalry, the AI Convergence, and the European Union's Strategic Role 53

Fabrizio d'Amore, Who is Behind the IP? Notes on an Unresolved Complexity 67

Paolo Dal Checco, Matteo Vinci, Marco Musumeci, Turning Digital Traces into Reliable Evidence 87

Andrea Accetta, From CVEs to Cyberspace Situational Understanding: Rethinking Vulnerability Intelligence in the Age of Artificial Intelligence 95

Ágnes Kasper, The Invisible Infrastructure of Cybersecurity Governance and the Role of the Cooperative Cyber Defence Centre of Excellence (CCDCOE) 115

Giorgio Marrapodi, Cyberspace and the United Nations: Towards a Global Framework for Responsible State Behaviour 131

Diego Brasioli, Cyberdiplomacy: Norms of Responsible State Behavior and the International Cooperation Framework 141

Annita Larissa Sciacovelli, Agentic Artificial Intelligence, Cybersecurity, and Governance, Risk, and Compliance (GRC): Challenges within the European and International Legal Frameworks 161

Cristiano Busco, *Mattia Siciliano*, Cyber Threat Intelligence, AI Agents and Cognitive Warfare: Elements of Critical Infrastructure 183

Louis Marinos, Ingredients of Risk Assessment Methods and Data to be Collected 191

Carmelo Dimauro, Redefining Cybersecurity Governance Against Adaptive Threats: From Classical Risk Analysis to Autonomous Cyber-Immunity 219

Corrado Aaron Visaggio, Cybersecurity for Artificial Intelligence 237

Mario Beccia, The Five Layers of AI 245

Cristine Skropke, Florian Sonnenschein, Torsten Redlich, Cybersecurity for Connected Medical Care: Lessons from the CYMEDSEC Research Project on Internet of Medical Things 265

Vita Santa Barletta, *Danilo Caivano*, Cyber-Social Security 275

Dominique Kogue, At the Heart of a Cyber Crisis: The Key to Success 295

Danielle Yeow, Singapore's National Cybersecurity Strategy and Resilience Model 315

Stefano Melina, Maritime Cybersecurity and Resilience: From NIS2 Compliance to the Protection of Critical Maritime and Underwater Infrastructure 339

Matteo Macina, Telecommunications: Securing the Backbone of the Digital Society and National Resilience 365

Daniele Frasca, *Vittoria Piantelli*, Space Systems as Critical Infrastructures: Cybersecurity Challenges for European Resilience and Strategic Sovereignty 381

FOREWORD

Cybersecurity has become a fundamental enabler of Europe's digital society, economy, and strategic autonomy. As digital technologies evolve at an unprecedented pace, our threat landscape shifts with them. Innovations in artificial intelligence, cloud computing, and quantum readiness offer immense opportunities, but they also introduce complex dependencies and vulnerabilities across our global supply chains. In this fast-paced environment, organisations face a double challenge: navigating a sophisticated threat landscape while complying with an evolving regulatory framework, including the NIS 2 Directive and the Cyber Resilience Act.

In addition to the technological and regulatory challenge of the organisations, governments are exposed, more and more to the effects of cyberattacks targeting critical infrastructures and public services. The very recent cases of cyber-attacks to supermarket chains in The Netherland and France show the significant impact that these attacks have on citizens and industry. In this regard, Critical infrastructures are used, by enemy states, as tool of geopolitical pressure.

The European Cybersecurity Industrial, Technology and Research Competence Centre (ECCC), alongside the Network of National Coordination Centres (NCCs) and the Cybersecurity Competence Community, were created with the goal of strengthen Europe's cyber resilience, of foster cross-border cooperation, and build a robust, self-reliant digital ecosystem across the Union. Achieving this requires bridging critical gaps: connecting research with market deployment, aligning policy with operational realities, and linking European initiatives with the concrete experiences of practitioners on the ground.

This Playbook is designed as a direct catalyst for that effort. Elaborated within the framework of the Jean Monnet European education programme, this volume embodies the core European values of academic excellence, knowledge dissemination, and intercultural cooperation. By embedding this initiative within Europe's premier educational framework, we ensure that cutting-edge cybersecurity expertise is not siloed, but systematically translated into an open, educational asset capable of training the next generation of European cyber leaders and practitioners.

For this reason, the Playbook is an essential tool to understand the complexity of the current cybersecurity ecosystem and how only a coordinated and multidisciplinary response can be successful in building a resilient and cyber-secure society.

The final scope of this Playbook is uniquely ambitious, reflecting a multidimensional approach to security with a heavy emphasis on the formulation and evolution of comprehensive cybersecurity strategies. It synthesises a vast array of specialised contributions that span the entire spectrum of modern cyberspace. Readers will find deep strategic insights into cyber-diplomacy, geopolitical dynamics, and the UN frameworks for responsible state behaviour. Crucially, the volume provides a rigorous comparative analysis of national cybersecurity strategies and governance models, evaluating both European frameworks and global benchmarks, such as Singapore's resilience model, to illustrate how structured public-private partnerships can successfully operationalise national resilience.

Furthermore, the Playbook bridges these high-level strategic definitions with deep operational and technical realities. It addresses the legal and operational friction points of Agentic AI, the technical frontlines of digital forensics, cyber threat intelligence, and vulnerability management (CVEs), and the defence of critical domains such as telecommunications, space, and maritime ecosystems. Crucially, it highlights the human element, drawing on pioneering operational practices like Luxembourg's ROOM#42 model, proving that strategy must always be grounded in practical execution and community engagement.

Rather than establishing new regulatory burdens or replacing existing standards, the purpose of this Playbook is to democratise this vast pool of collective intelligence. It provides students, researchers, policy-makers, and industry operators with a comprehensive, educational, and practical compass to design robust strategies and navigate the complexities of our shared digital space.

In a domain as dynamic as cybersecurity, continuous learning, strategic foresight, and structured knowledge-sharing are our strongest defences. I trust that this Playbook, enriched by the spirit of the Jean Monnet programme, will serve as a definitive source of inspiration, education, and practical guidance for everyone dedicated to safeguarding Europe's digital future.

Luca Tagliaretti
Executive Director
European Cybersecurity Competence Centre, Romania

PREFACE OF THE EDITORS

This open-access book is funded by the European Commission within the framework of the project activities and it contributes to the dissemination activities of the Jean Monnet Module "Cybersecurity in the Digital Era: European Perspectives," implemented by the University of Bari Aldo Moro. Developed in collaboration with the Luxembourg House of Cybersecurity (LHC), it brings together legal, strategic, and operational perspectives on cybersecurity and artificial intelligence, reflecting their growing importance for European governance and the digital transformation of society.

In this context, the Playbook enhances to the broader objectives of the Jean Monnet Programme by promoting knowledge and understanding of European approaches to cybersecurity and digital transformation. It also encourages dialogue across academic, institutional, and professional communities, and connects European policy and regulatory perspectives with international developments and practical experience. By bringing together different disciplines and perspectives, the volume aims to support informed reflection on how European values, institutions, policies, and legal frameworks shape responses to emerging challenges in the digital domain.

Cybersecurity has become a condition for the exercise of public authority, the continuity of essential services, and the protection of individual rights. Decisions about how digital systems are secured affect the ability of institutions to govern, of businesses to operate, and of people to participate in democracies. Artificial intelligence deepens these dependencies. It can strengthen the detection and analysis of threats while introducing new vulnerabilities into systems that increasingly generate content, support decisions, and act across connected environments. Understanding this relationship requires an account of how strategic choices, legal duties, operational practices, and human capabilities work together.

This Playbook brings those perspectives into a common conversation. Its chapters draw on legal scholarship, public policy, diplomacy, law enforcement, technical research, and operational experience. Each field asks different questions, yet they depend on shared foundations: reliable information, clear responsibilities, cooperation across borders, and the capacity to respond when preventive measures fail. We have written this volume to make those connections visible and usable. It moves from the frameworks that allocate authority and shape strategy, through international cooperation and the management of emerging threats, to the resilience of people, organisations, and critical sectors.

It's first part establishes the legal and strategic setting. It begins with the institutional architecture of cybersecurity and the regulatory frameworks through which

responsibilities are distributed. This is the necessary starting point for assessing governance: who has authority to act, what duties apply, and how European and national levels of decision-making interact. The discussion then turns to national cybersecurity strategies, sovereignty, and public-private partnerships.

The analysis next widens to the geopolitical contest surrounding cyberspace and artificial intelligence. European choices take shape amid competition over technological capabilities and infrastructure. A comparative account of Singapore's cybersecurity strategy adds another perspective on how a state can organise preparedness and resilience.

The second part of the Playbook focuses on how States and institutions cooperate when cyber activity crosses borders. It examines the United Nations' work on responsible State behaviour in cyberspace, the role of cyber diplomacy in developing norms and sustaining cooperation, and the contribution of specialised international institutions to research, training, and shared expertise. These chapters place European cybersecurity governance within a wider international setting. They also show that agreement on broad principles is only one stage of cooperation. States and other actors must continue to address differences of interpretation, build practical capabilities, and find ways to apply shared commitments under changing technological conditions.

The third part of the Playbook turns from frameworks to the changing threat environment. Cyber threat intelligence provides information about adversaries, vulnerabilities, and methods of attack; risk analysis helps organisations assess what that information means for their own systems and responsibilities. Together, these disciplines provide an operational foundation for the chapters on artificial intelligence. The volume considers the security of AI systems, the governance of agentic AI, AI's relevance to strategic autonomy, and the evolution from information about individual vulnerabilities towards wider a cyberspace situational awareness.

The final part of the Playbook examines what resilience means where cyber risk is experienced most directly. It begins with the social dimensions of digital security and the role of immersive simulation in preparing people and organisations to act under pressure. It then moves across connected systems in industry and healthcare, maritime and port operations, telecommunications, and space. Each setting has its own technologies, dependencies, and regulatory context. All these chapters give concrete form to the strategic and legal issues developed earlier in the volume.

The book thus follows a deliberate sequence: from authority, to cooperation, to risk, and finally to resilience. Each stage informs the next. Legal obligations need institutions capable of carrying them out; international cooperation needs both shared understandings and reliable evidence; risk analysis needs knowledge of the systems and people exposed; and resilience needs preparation that can be tested in operational settings. No single perspective can answer all these questions. The value of the Playbook lies in showing how they meet, and where gaps between them still require attention.

This progression reflects a central concern of the book: we are entering a transition phase in which artificial intelligence is fundamentally reshaping the cybersecurity landscape. Understanding the legal implications of these developments requires more than knowledge of legal rules alone. Students and practitioners need to understand what is at stake: the technologies being developed and deployed, the strategies and policies that shape their use, the risks and opportunities they create, and the operational environments in which decisions are made.

An autonomous AI system that assists analysis may itself depend on data, models, tools, and external connections that require scrutiny. It's capable of carrying out a sequence of actions that may accelerate a legitimate process while raising difficult questions about oversight, accountability, and the effectiveness of human intervention. These realities provide the context in which questions of legal responsibility, governance, compliance, and accountability can be properly understood. Technical assessment, organisational governance, strategic choices, and legal responsibility must therefore be considered together.

As editors, we hope this volume will serve readers who work across those boundaries: scholars and students, policymakers, legal practitioners, investigators, security professionals, and operators of essential services. Bringing these readers together is itself part of the book's purpose, as effective cybersecurity governance depends on their ability to understand one another's responsibilities and work together.

Last but not least, we express our deepest gratitude to all the contributing authors. We are profoundly thankful to the distinguished experts from NATO, European Union, the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE), EUROPOL, United Nations, members of the diplomatic community, pioneering technology companies, critical infrastructure operators, and colleagues from esteemed academic institutions worldwide.

Their generosity in sharing specialized knowledge, their dedication to intellectual rigor, and their willingness to bridge the worlds of policy, diplomacy, law enforcement, corporate enterprise, and critical operations made this volume possible. Their contributions reflect a shared commitment to building a secure, open, resilient, and rights-respecting digital society.

We also convey our sincere thanks to the European Commission and the Jean Monnet Programme, whose financial support enabled this open-access initiative.

Annita Larissa Sciacovelli
University of Bari Aldo Moro, Italy
Carmelo Dimauro
Luxembourg House of Cybersecurity, Luxembourg
Mattia Siciliano
Luiss Business School, Italy

THE NORMATIVE AND INSTITUTIONAL ARCHITECTURE OF EU CYBERSECURITY AND CYBER DEFENCE: BETWEEN MARKET INTEGRATION AND NATIONAL SOVEREIGNTY

Annita Larissa Sciacovelli
University of Bari Aldo Moro, Italy

Abstract
The European Union has built a multi-layered, decentralized system to govern cyberspace, balancing national sovereignty with collective security. In this architecture, national governments maintain primary responsibility for day-to-day oversight, emergency response, and national defence, while EU institutions provide common rules, political coordination, pooled capabilities, and shared Cyber Threat Intelligence (CTI). This chapter explores the legal groundings and practical mandates of central EU actors, including the European Commission, the EU Agency for Cybersecurity (ENISA), the NIS Cooperation Group, the CSIRTs network, the European Cyber Crises Liaison Organisation Network (EU-CyCLONe), and the European Cybersecurity Competence Centre (ECCC). It examines how EU laws in the cybersecurity sector – particularly Directive (EU) 2022/2555 (NIS2 Directive), Regulation (EU) 2019/881 (Cybersecurity Act), Regulation (EU) 2024/2847 (Cyber Resilience Act), and Regulation (EU) 2025/38 (Cyber Solidarity Act) – interact with specialized regimes across the digital single market. Crucially, the chapter integrates Regulation (EU) 2024/1689 (Artificial Intelligence Act) and further investigates how civilian cybersecurity policies converge with joint military cyber defence initiatives.

1. Introduction: The Distributed Governance in a Complicated Cyber Threat Landscape

Cybersecurity has grown into an expansive, transversal regulatory domain in European Union (EU) law, and it is situated at the intricate structural junction of internal-market integration, public order maintenance, consumer and product safety, critical infrastructure resilience, and the Common Foreign and Security Policy. It is well known that a single cross-border cyber incident – such as a sophisticated supply-chain intrusion compromising an enterprise cloud hosting environment that simultaneously supports regional hospital networks, municipal water distribution, cross-border financial transactions, and public administration portals – readily demonstrates the profound legal and institutional friction points of this regulatory ecosystem.

An event of this magnitude immediately and concurrently activates operational incident containment protocols, mandatory personal data breach

notifications under Regulation (EU) 2016/679 (General Data Protection Regulation), market surveillance evaluations concerning vulnerable commercial hardware and software components, domestic criminal investigations into unlawful network interference, and high-level foreign-policy deliberations regarding diplomatic attribution. Because multiple legally protected societal interests could collide instantaneously within a single significant technical disruption, an extensive, pluralistic array of public authorities is compelled to respond simultaneously. The central legal problem confronting the European Union is determining the precise legal conditions, authorities, and procedural mechanisms through which competent EU bodies and domestic authorities can lawfully share threat telemetry, coordinate technical containment measures, and exercise public regulatory authority across national borders without eroding the legal boundaries established by primary EU law.

To manage these compounding vulnerabilities, the EU has deliberately chosen a *model of functionally distributed, polycentric governance* rather than attempting to establish an all-encompassing, centralized supranational cyber authority endowed with direct control over domestic networks. European institutions, specialized agencies, and cooperation networks exist primarily to supplement, support, and harmonize national authorities without displacing the sovereign regulatory, administrative, and operational competences of the Member States across four basic functions of this model.

Specifically this model works first thanks to the EU normative harmonisation that establishes binding, baseline security and risk-management requirements for economic operators, public administrations, and digital products across the single market, preventing market fragmentation. Second, the model is based on prevention, standardization, and knowledge dissemination that are coordinated at the Union level by the European Union Agency for Cybersecurity-ENISA through technical certification frameworks, coordinated vulnerability disclosure repositories, and systematic threat analysis that are operating in conjunction with domestic market surveillance authorities. Third, the model has established an incident handling and crisis management that follow a structured operational escalation ladder that runs from domestic computer security incident response teams (CSIRT), through the cross-border Computer Security Incident Response Teams (CSIRTs) Network, to the European Cyber Crisis Liaison Organisation Network (EU-CyCLONe) and high-level political crisis channels. Fourth, all the previous functions are based on technological capability development that is driven by the European Cybersecurity Competence Centre-ECCC alongside the collective emergency mechanisms established by the Cyber Solidarity Act.

In this domain it is worth noting that the regulatory enforcement and the

exercise of coercive public power remain the exclusive prerogative of domestic authorities, sector-specific regulators, and national judicial bodies acting in strict accordance with national procedural codes.

2. The Delimitation of the European Union Competences

The legal basis of the EU cybersecurity legislative body is Article 114 of the Treaty on the Functioning of the European Union (TFEU), which empowers the European Parliament and the EU Council to adopt harmonising measures for the establishment and functioning of the internal market. The European legislature relies predominantly on this Article to harmonize regulatory standards that are necessary to guarantee a uniform, high level of cybersecurity across all Member States. However, the recourse to internal-market competence is strictly circumscribed by the principle of conferral under Article 5(2) of the Treaty on European Union (TEU) and the sovereign reservation of national security embodied in Article 4(2) TEU, which states categorically that national security remains the sole responsibility of each individual Member State.

This fundamental division of competences directly governs daily operational collaboration, cross-border threat intelligence exchanges, and administrative coordination within the EU networks. Whenever technical indicators of compromise, vulnerability reports, or incident notifications are transmitted through European cooperation platforms, institutional interoperability requires unyielding adherence to statutory purpose limitation and domestic procedural safeguards. The technical submission of network telemetry to a national CSIRT or its subsequent transmission across the CSIRTs Network does not determine whether that data can be lawfully admitted as evidence within subsequent domestic criminal proceedings, nor does it establish a legally binding attribution of State responsibility under public international law. Preserving the rule of law within this distributed architecture requires maintaining clear, uncompromised boundaries between rapid technical containment, judicial accountability, and sovereign diplomatic attribution, ensuring that regulatory reporting mechanisms established to foster internal-market resilience are not quietly converted into instruments of warrantless intelligence gathering, public surveillance, or domestic policing outside the established constitutional framework.

3. The Cyber Threat Intelligence and the Cooperative Core of the NIS2 Framework

Within this legal structure, Cyber Threat Intelligence (CTI) – encompass-

ing the structured collection, forensic analysis, contextualization, and automated exchange of information regarding threat actor profiles, tactics, techniques, and procedures (TTPs), and technical indicators of compromise – serves as the indispensable operational element of the EU entire defensive network. Rather than operating as an optional technical diagnostic, real-time threat intelligence provides the practical and legal precondition for proactive risk mitigation under Article 21 of the Directive (EU) 2022/2555 on Network and Information Security (NIS2 Directive), early warning notifications under Article 23, and coordinated vulnerability disclosure routines under Article 12. Without access to high-fidelity, actionable threat intelligence, statutory duties of organizational resilience remain purely theoretical, as regulated entities cannot realistically anticipate zero-day vulnerabilities, detect sophisticated living-off-the-land techniques, or configure network firewalls against State-sponsored intrusion sets operating at machine speed.

This operational reality is anchored by ENISA under Regulation (EU) 2019/881 (Cybersecurity Act), which permanently establishes the Agency as the central technical analysis and knowledge hub for the Union, charging it with assisting the European Commission and Member States, coordinating certification frameworks, and administering the secretariats of operational cooperation bodies.

The empirical findings published in the ENISA Threat Landscape 2026 report underscore the systemic necessity of this mandate: nearly three-quarters of all major cyber incidents across the European Union now directly target essential and important entities covered under the NIS2 Directive. The public administration sector remains the primary target, absorbing nearly a third of all documented incidents, heavily disrupted by Distributed Denial of Service (DDoS) campaigns, while software supply-chain exploits and the weaponization of artificial intelligence models by State-backed and criminal intrusion groups continue to accelerate. To convert this intelligence into defensive coordination, the NIS2 Directive establishes three distinct administrative layers: the NIS Cooperation Group, under Article 14, that handles strategic policy the CSIRTs Network, under Article 15, that provides an operational forum for technical exchange and joint mitigation among national teams and CERT-EU, and the EU-CyCLONe, under Article 16, that coordinates executive crisis management during large-scale incidents, feeding contextualized assessments directly into the EU Council's Integrated Political Crisis Response arrangements. Although domestic authorities handle audits and direct supervision, operational escalation across these European networks boosts collective situational awareness.

4. Building Cyber Capabilities, Product Security, and Sectoral Resilience

The EU supplements the administrative reporting and the operational coordination with industrial cyber capability development, mandatory product lifecycle standards, and specialized sectoral regimes also to eliminate strategic technological dependencies. In depth, industrial capability development is directed by the European Cybersecurity Competence Centre (ECCC) in Bucharest, established by Regulation (EU) 2021/887, working in partnership with the Network of National Coordination Centres. The ECCC pools financial resources from the Digital Europe and Horizon Europe programmes to enhance industrial capacity, advance dual-use technological research, and foster European cybersecurity skills, thereby linking normative regulatory demands with the material technological capacity required for market implementation.

At the horizontal product level, Regulation (EU) 2024/2847 (Cyber Resilience Act) introduces comprehensive security duties for all hardware and software products with digital elements brought to the EU market. The Cyber Resilience Act mandates cybersecurity-by-design principles throughout the entire product lifecycle, obliges manufacturers to implement systematic vulnerability management, and introduces mandatory reporting channels requiring manufacturers to notify ENISA and designated national CSIRTs of actively exploited vulnerabilities within strict statutory deadlines. Concurrently, the European cybersecurity certification framework established by the Cybersecurity Act – exemplified by the European Common Criteria-based scheme (EUCC) adopted via the European Commission Implementing Regulation (EU) 2024/482 – provides formal conformity assessments that attest to defined security benchmarks, although holding a valid certificate creates only a rebuttable presumption of conformity and does not grant immunity from independent liabilities under the NIS2 Directive or data protection law.

The broad baseline set by the NIS2 Directive interfaces with targeted sectoral frameworks that take precedence in specialized fields, most notably Regulation (EU) 2022/2554 on Digital Operational Resilience for the Financial Sector (DORA) and Directive (EU) 2022/2557 on the resilience of critical entities (CER Directive). Crucially, this regulatory architecture also incorporates Regulation (EU) 2024/1689 (Artificial Intelligence Act), which establishes a comprehensive fundamental-rights regime governing algorithmic models. Under Article 15 of the EU AI Act, high-risk artificial intelligence systems – including algorithmic components integrated into critical digital infrastructures – must be designed and developed to achieve an appropriate level of technical robustness, accuracy, and cybersecurity throughout their entire lifecycle. High-risk systems must be resilient against attempts by unauthorized third parties to alter their performance or exploit vulnerabilities, specifically guarding against

AI-specific attack vectors such as data poisoning, model poisoning, adversarial inputs, and model inversion breaches.

The inclusion of the EU AI Act within the broader cybersecurity framework is indispensable for addressing the emerging governance challenges especially those posed by Agentic Artificial Intelligence Systems (AAI). Unlike conventional machine learning models that generate static predictions or content, AAI systems operate with substantive cognitive autonomy, possessing the capability to reason, orchestrate workflows, make independent decisions, and execute multi-step operational actions across networks without continuous human oversight. While this functional autonomy offers immense defensive advantages in monitoring threat telemetry and automating real-time incident containment at machine speed, it simultaneously introduces acute systemic vulnerabilities. Malicious actors can co-opt autonomous agents to conduct millisecond-scale polymorphic attacks, exploit semantic feedback loops, or manipulate decision logic. By mandating technical fail-safes, human oversight architectures, and lifecycle cyber resilience under Article 15, the EU AI Act provides the essential statutory baseline upon which subsequent legal analysis of autonomous agentic systems must build.

5. Crisis Assistance, Solidarity, and the Need for a Cyber Civil Defence Mindset

To provide concrete mutual aid during severe transboundary emergencies, Regulation (EU) 2025/38 (Cyber Solidarity Act) introduces a structured mutual assistance mechanism organized across three pillars: the European Cybersecurity Alert System, the Cybersecurity Emergency Mechanism, and the European Cybersecurity Incident Review Mechanism. The Alert System interconnects national and cross-border cyber hubs across the Union to pool threat telemetry, improve early detection, and provide real-time situational awareness. The Emergency Mechanism provides targeted preparedness testing for critical sectors alongside access to an EU Cybersecurity Reserve of private, pre-vetted incident response providers deployable to assist affected Member States, while the *Review Mechanism* tasks ENISA and EU-CyCLONe with conducting post-incident reviews of major disruptions to derive regulatory and systemic lessons for the internal market. Support from the Reserve remains subject to the voluntary request of an affected Member State demonstrating capacity exhaustion, ensuring that the Regulation creates assistance options without granting Union bodies authority to commandeer national networks.

In our opinion, this shift toward collective operational support requires a

deeper transformation in strategic thought regarding critical infrastructure protection. As articulated in the study of 2025 on *Creating a Civil Defense Mindset to Address Modern Cyber Threats* of the MITRE Corporation, modern cyber threats targeting civilian essential services require a fundamental transition from traditional, siloed network defence to an updated *civil defence mindset* because in the contemporary digital domain, critical services – power distribution, telecommunications, water utilities, logistics, and healthcare – are overwhelmingly owned and operated by private entities that operate on the *frontlines* of state-sponsored offensive campaigns. Relying solely on conventional governmental response teams or sporadic regulatory audits is inadequate when transboundary attacks can paralyze societal functions in minutes.

Adapting this civil defence model to the European Union requires treating private infrastructure operators not merely as *passive* recipients of regulatory mandates under NIS2 or the CER Directives, but as frontline operational participants in collective defence. As the MITRE Study emphasize, an effective civil defence posture requires structured, pre-established operational integration: collective incident playbooks, continuous civil-military telemetry and cyber threat intelligence sharing, pre-vetted private auxiliary capabilities, and widespread societal cyber hygiene and continuity training. The Cyber Solidarity Act's private emergency reserve and Early Warning Alert System represent initial legislative movements toward this philosophy. However, operationalizing a genuine civil defence posture in Europe demands reconciling this collective mobilization with strict legal boundaries, ensuring that private operators assisting in emergency response possess clear legal protections, distinct mandates, and defined lines of public accountability.

6. Strategic Visions: Bridging Civilian Resilience and Military Cyber Defence

Actually, the EU overarching strategic direction is guided by high-level policy roadmaps that span different chapters of the EU Treaties, notably the Joint Communication on The *EU's Cybersecurity Strategy for the Digital Decade* (2020) and the Joint Communication on an *EU Policy on Cyber Defence* (2022). Adopted by the European Commission and the High Representative, the civilian Cybersecurity Strategy outlines three interdependent policy goals: building resilience, technological sovereignty, and leadership; developing operational capacity to prevent, deter, and respond; and promoting an open, global cyberspace through international partnerships. As a joint communication, the Strategy represents an agenda-setting soft-law document that neither creates directly enforceable obligations for private operators nor confers new statutory powers upon EU agencies, deriving its legal importance from articulating the

political objectives that later rematerialized in the NIS2 Directive, the Cyber Resilience Act, and the Cyber Solidarity Act.

In parallel, the EU Policy on Cyber Defence was adopted following the EU Council's approval of the *Strategic Compass for Security and Defence* (2022), articulating a collaborative defence posture organized across four pillars: acting together through closer coordination between national military response teams and civilian bodies; securing the Union's defence ecosystem against espionage and supply-chain compromise; investing in sovereign military cyber capabilities through the European Defence Fund and Permanent Structured Cooperation; and partnering with third states and NATO while preserving Union decision-making autonomy. Like the civilian strategy, this policy document does not transfer sovereign military command to the Union, leaving operational cyber command, control, and offensive capabilities entirely under Member State authority. This capability dimension was subsequently expanded by the White Paper for European Defence – Readiness 2030 (2025) and the Joint Communication on the Defence Readiness Roadmap 2030 (2025), which designate cyber capabilities, artificial intelligence, and electronic warfare as critical capability domains for joint investment.

7. The Legal Boundary Between Civilian Cybersecurity Defence and Cyber Military Action

The operational relationship between civilian cybersecurity and cyber defence is complementary because military command systems and operational mobility rely heavily on commercial telecommunications and cloud infrastructures, and civilian operators often need military intelligence that identifies advanced persistent threat groups targeting civilian critical infrastructure. However, this operational interdependence does not dissolve the primary Treaty boundaries separating the internal market from the UE Common Security and Defence Policy (CSDP).

Maintaining legal coherence requires enforcing procedural safeguards across several intersections. First, information collected under NIS2 Directive supervisory mechanisms cannot be automatically transferred into classified military repositories without satisfying strict purpose-limitation rules, data protection principles, and statutory domestic guarantees. Second, a malicious cyber operation against critical infrastructure does not automatically trigger the mutual assistance clause of Article 42(7) TEU or the solidarity clause of Article 222 TFEU, as both provisions require explicit political evaluation by the Member States confirming that the disruption meets the stringent legal thresholds of armed aggression or natural or man-made disaster. Third, legal classifications

for emerging technologies diverge across regimes; while the EU AI Act introduces harmonised safety standards for civilian deployments, systems developed exclusively for military, defence, or national-security purposes are expressly excluded from its scope under Article 2(3) of the Act, necessitating contextual legal analysis rather than the assumption of a single regulatory regime.

8. Conclusions

The fundamental inquiry animating this study is whether the European Union's normative and institutional architecture, divided between internal-market development under Article 114 TFEU and the sovereign reservation of national security under Article 4(2) TEU, can effectively safeguard European cyberspace against contemporary cyber threats. The analysis confirms that the traditional paradigm of siloed, passive administrative compliance is structurally inadequate. In an operational environment where State-sponsored advanced persistent threats target private cloud infrastructure, telecommunications backbones, and municipal utilities below the threshold of kinetic war, the division between civilian cybersecurity and national defence becomes functionally obsolete. As underscored by the empirical data of the ENISA Threat Landscape 2026 and the analytical insights MITRE Study, modern cyber conflict demands a comprehensive civil defence mindset. Essential societal services are overwhelmingly owned and operated by private commercial entities; consequently, societal survival depends on integrating these private operators, civilian response networks, and military cyber defence bodies into a coordinated, collective posture. The central legal challenge is translating this strategic necessity into legally sound mechanisms that uphold the rule of law, respect the principle of conferral under Article 5(2) TEU, and protect fundamental rights.

Achieving this objective requires establishing concrete legal conduits for civil-military collaboration, with particular emphasis on Cyber Threat Intelligence (CTI) sharing governed by statutory *procedural firewalls*. In legal and institutional terms, procedural firewalls are legally mandated information barriers and procedural safeguards designed to restrict, filter, and control the flow of data across different jurisdictions, mandates, and competence regimes. Unlike technical firewalls that filter digital packets, a procedural firewall operates as a normative barrier that prevents the unlawful conversion or secondary use of data gathered under civilian internal-market regulations into classified military intelligence or criminal proceedings. Under Directive NIS2 and Cyber Resilience Act, commercial entities are subjected to mandatory incident and vulnerability notification duties toward civilian CSIRTs. If that proprietary threat telemetry were permitted to flow freely into the EU Common Security and

Defence Policy structures, military emergency response teams (milCERTs), or domestic intelligence repositories without strict legal separation, it would generate a severe chilling effect, discouraging private infrastructure operators from sharing threat data due to the fear of regulatory self-incrimination, commercial exposure, or unlawful intelligence surveillance.

To resolve this operational and legal problem, the European legislator should enact an *Inter-Institutional Information-Sharing Protocol* that formalizes these procedural firewalls within statutory text. Such an instrument should integrate four distinct legal safeguards. First, it should enforce the principle of purpose limitation by establishing that telemetry submitted by private entities for defensive containment may be processed solely to mitigate the immediate technical incident, prohibiting its secondary use for market sanctions or general surveillance. Second, it should introduce statutory safe harbours and evidentiary protections, ensuring that good-faith technical disclosures made during an ongoing crisis cannot be admitted as evidence of corporate fault in subsequent civil litigation or domestic regulatory enforcement. Third, it should mandate automated technical sanitization protocols, requiring civilian authorities to strip proprietary business data, personal information, and identifiable architecture logs from threat reports before transferring them to military or foreign-policy bodies, converting raw commercial data into generic, non-attributable indicators of compromise. Fourth, it should ensure that identifiable domestic network logs can be accessed by military intelligence or law enforcement only pursuant to an independent judicial authorization under national law. By establishing these statutory firewalls, the Protocol would reconcile machine-speed intelligence sharing with the fundamental rights guaranteed by the Treaties.

To complement this intelligence-sharing framework and fully operationalize the civil defence paradigm, the Union must implement three additional structural reforms. First, the European Commission and ENISA must establish a Codified Standard Operating Procedure and Civil Defence Auxiliary Charter under the Cyber Solidarity Act, setting down clear cross-border liability immunities, emergency rules of engagement, and statutory thresholds of domestic capacity exhaustion for private incident responders deployed via the EU Cybersecurity Reserve. Second, the European legislature should introduce a dedicated Supplementary Directive on Autonomous and Agentic Systems in Critical Infrastructures, linking Article 15 of the EU AI Act with Article 21 NIS2 to mandate cryptographic execution logging and automated machine-to-machine validation protocols for autonomous defence agents. Third, the Council should codify an institutionalized coordination channel between EU-CyCLONe and the EU Political and Security Committee, guaranteeing high-confidence evidentiary scrutiny before operational threat telemetry is transitioned into diplomatic countermeasures under the Cyber Diplomacy Toolbox or mutual defence

under Article 42(7) TEU. Through these targeted legal mechanisms, the European Union can transcend the artificial divide between market regulation and sovereign defence, establishing a resilient cyber posture anchored in collective solidarity, operational efficacy, and the primary law of the Treaties.

References

Consolidated Version of the Treaty on European Union, OJ C 202, 7.6.2016.

Consolidated Version of the Treaty on the Functioning of the European Union, OJ C 202, 7.6.2016, p. 47.

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 (General Data Protection Regulation), OJ L 119, 4.5.2016, p. 1.

Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification (Cybersecurity Act), OJ L 151, 7.6.2019, p. 15.

Regulation (EU) 2021/887 of the European Parliament and of the Council of 20 May 2021 establishing the European Cybersecurity Industrial, Technology and Research Competence Centre and the Network of National Coordination Centres, OJ L 202, 8.6.2021, p. 1.

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 (NIS2 Directive), OJ L 333, 27.12.2022, p. 80.

Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of critical entities (CER Directive), OJ L 333, 27.12.2022, p. 164.

Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (DORA), OJ L 333, 27.12.2022, p. 1.

Commission Implementing Regulation (EU) 2024/482 of 31 January 2024 laying down rules for the application of Regulation (EU) 2019/881 as regards the adoption of the European Common Criteria-based cybersecurity certification scheme (EUCC), OJ L 2024/482, 7.2.2024.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L 2024/1689, 12.7.2024.

Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act), OJ L 2024/2847, 20.11.2024.

Regulation (EU) 2025/38 of the European Parliament and of the Council of 15 January 2025 laying down measures to strengthen solidarity and capacities in the Union to detect, prepare for and respond to cybersecurity threats and incidents (Cyber Solidarity Act), OJ L 2025/38, 22.1.2025.

Council of the European Union, Council Conclusions on a Framework for a Joint EU Diplomatic Response to Malicious Cyber Activities (Cyber Diplomacy Toolbox), Doc. 9916/17, 19.6.2017.

European Commission and High Representative of the Union for Foreign Affairs and Security Policy, Joint Communication: The EU's Cybersecurity Strategy for the Digital Decade, JOIN(2020) 18 final, 16.12.2020.

Council of the European Union, A Strategic Compass for Security and Defence – For a European Union that protects its citizens, values and freedoms and contributes to international peace and security, Doc. 7371/22.3.2022.

European Commission and High Representative of the Union for Foreign Affairs and Security Policy, Joint Communication: EU Policy on Cyber Defence, JOIN(2022) 49 final, 10.11.2022.

Council of the European Union, Council Conclusions on the EU Policy on Cyber Defence, Doc. 9618/23.5.2023.

Council of the European Union, Council Conclusions on the Future of EU Cybersecurity, Doc. 10133/24.5.2024.

Council of the European Union, Council Recommendation on an EU Blueprint for Cyber Crisis Management, C/2025/3445, 6.6.2025.

European Commission and High Representative of the Union for Foreign Affairs and Security Policy, Joint Communication: Preserving Peace – Defence Readiness Roadmap 2030, JOIN(2025) 27 final, 16.10.2025.

European Union Agency for Cybersecurity (ENISA), ENISA Threat Landscape 2026: Exploring the Evolution of the Cyber Threat Landscape and Systemic Dependencies, 22.9.2026.

Bristow, Mark, and Irving Lachow, Past is Prologue: Creating a Civil Defense Mindset to Address Modern Cyber Threats, The MITRE Corporation, McLean, VA, May 2025.

CYBERCRIME & LEA

Emmanuel Kessler
Europol – O3 European Cyber Crime Centre (EC3), The Netherlands

Abstract
The provided text outlines the complex evolution of the global cybercrime ecosystem, emphasizing how isolated illicit platforms have transformed into interconnected networks driven by dark-web services, artificial intelligence, and cryptocurrencies. Within the European Union, severe threats like ransomware, online fraud, and a blurring boundary between cybercrime and hybrid operations demand robust cross-border responses and close public-private partnerships. Europol's European Cybercrime Centre (EC3) serves as a vital operational anchor, coordinating international investigations through mechanisms like the Joint Cybercrime Action Task Force. However, these enforcement capabilities remain heavily reliant on unified judicial frameworks – such as the Budapest Convention, the UN Convention against Cybercrime, and the EU's NIS2 Directive – to secure timely data access under strict privacy safeguards. Despite these cooperative and legislative steps forward, investigations continue to face significant challenges from anonymization technologies, data volatility, and divergent legal systems. Moving forward, law-enforcement bodies must continuously enhance their strategies to keep pace with emerging technical capabilities, particularly in the realms of artificial intelligence and post-quantum computing.

1. Cybercrime Threat Landscape

The fight against cybercrime is a key pillar of efforts to ensure the security of the digital environment, complementing the work of cybersecurity stakeholders. It aims to prevent and counter cybercrime, enhance the collection and exchange of information on cybercriminal actors and threats, protect victims, and disrupt the online criminal ecosystem.

Enhancing knowledge and understanding of the cybercrime ecosystem, including the identification of the criminal actors, is a fundamental prerequisite that shapes effective law enforcement action. The constant evolution of the ecosystem towards more complexity, as observed, highlights the importance of cooperation in the fight against cybercrime, and the need of a fast adaptation for law enforcement and cybersecurity actors.

This cybercrime ecosystem is analysed in depth in the 2026 edition of Europol's Internet Organized Crime Threat Assessment (Europol, 2026). Cybercrime continues to evolve towards greater fragmentation, specialisation and technological sophistication. On their side, criminal actors increasingly com-

bine dark-web services, encrypted communications, decentralised technologies, anonymisation infrastructure, cryptocurrencies and artificial intelligence (AI), to get anonymised whenever possible.

The dark web, which remains a central component of the cybercriminal ecosystem, offers anonymisation, although its structure is changing. Large general-purpose marketplaces are increasingly being replaced by smaller and more specialised platforms offering fraud services, compromised credentials, malware, access brokerage and other criminal capabilities. International law-enforcement operations have shortened the lifespan of major criminal markets and increased mistrust among users and administrators. However, disruption frequently results in rapid displacement rather than lasting disappearance.

Anonymisation of attackers also relies on infrastructures, as cybercriminals increasingly operate within a hybrid environment combining the clear web, dark web, end-to-end encrypted (E2EE) communications and decentralised services. Bulletproof hosting remains important, particularly when infrastructure is distributed across multiple providers and jurisdictions. VPN chaining, Tor routing and residential proxy networks add successive layers of anonymity, complicating attribution and generating additional requirements for cross-border cooperation. More sophisticated criminal organisations are also developing proprietary infrastructure, thereby reducing their dependence on commercial providers and limiting opportunities for law enforcement to identify or seize backend systems.

Anonymisation is also linked with the cryptocurrencies' challenge. Similarly, cryptocurrencies have evolved into a broader instrument for concealing proceeds across serious and organised crime. Ransomware actors continue to rely extensively on cryptocurrencies, increasingly using privacy-enhancing assets and exchange services located in jurisdictions with weaker anti-money-laundering controls. Blockchain bridges, decentralised exchanges and automated mixing services enable rapid movement between different blockchains and complicate transaction tracing. At the same time, wider public adoption of cryptocurrencies has expanded the potential victim population, including younger users exposed to investment fraud and wallet-draining schemes.

Domain Name System (DNS) infrastructure constitutes another significant enabler. Malicious domains connect criminal infrastructure with victims through phishing, malware distribution, credential theft and command-and-control activities.

Beyond anonymisation, cybercriminal actors increasingly exploit the opportunities offered by technological developments and innovation to enhance the scale, efficiency and resilience of their criminal activities.

AI and automation are becoming cross-cutting accelerators throughout the cybercrime ecosystem. Generative AI enables criminals to improve coding, so-

cial-engineering scripts, fraudulent communications and operational planning. Malicious or modified large language models are distributed through underground communities, while legitimate AI applications are also exploited for criminal purposes. These developments are particularly visible in online fraud schemes (OFS). During 2025, the threat evolved from one primarily characterised by volume and diversity towards one increasingly defined by speed, automation and sophisticated operational security. Very large online platforms facilitate mass victim targeting while reducing linguistic and geographical barriers.

The use of telecommunications infrastructure has also become increasingly important. Industrial-scale SIM farms allow criminal networks to operate thousands of accounts, circumvent identification requirements and localise fraudulent communications to specific countries. International Mobile Subscriber Identity (IMSI) catchers and SMS-blasting technologies provide additional mechanisms for bypassing legitimate telecommunications networks and disrupting traditional investigative evidence chains. Virtual desktop services, caller-ID spoofing and rapidly created neo-bank accounts further reinforce this criminal infrastructure. Payment fraud has also evolved technically, but despite technological advances, human vulnerability remains fundamental. Phishing and smishing have adapted to the wider implementation of two-factor and biometric authentication, while social engineering continues to provide mechanisms for circumventing technical safeguards.

Ransomware remains one of the most significant cyber threats affecting the EU. The ecosystem is increasingly fragmented, with more than 120 ransomware brands observed during 2025 and considerable overlap between operators, affiliates, infrastructure and money-laundering networks. Leaked ransomware source code, Ransomware-as-a-Service (RaaS) platforms and AI-assisted development tools have reduced technical barriers and enabled criminal actors to create or rapidly rebrand ransomware operations. Initial access continues to rely on phishing, infostealers, exploitation of vulnerabilities, Initial Access Brokers and insider threats. The ransomware business model has nevertheless shifted significantly, with extortion increasingly focusing on preventing the publication of stolen information rather than simply restoring encrypted data. As organisations improve backup and recovery capabilities, the reputational, regulatory and commercial consequences of data disclosure provide criminals with stronger leverage. The ransomware landscape encompasses RaaS programmes, more selective semi-closed structures and sophisticated closed groups. Platforms increasingly provide integrated criminal ecosystems covering malware deployment, persistence, exfiltration, negotiation, leak-site hosting and laundering services. Closed groups remain particularly challenging because they rely less extensively on commercial criminal services, operate proprietary infrastructure and maintain stronger internal trust and operational security.

The boundaries between conventional cybercrime, hybrid threats and other online criminal communities are becoming increasingly blurred. Cybercriminal infrastructure can be used by state-linked actors for distributed-denial of service (DDoS) attacks, intrusions, ransomware or information theft, while criminal actors may knowingly or unknowingly act as proxies. Emerging coalitions between established hacking communities further illustrate the fluidity of this environment. DDoS attacks remain particularly attractive because inexpensive and widely available services enable sustained disruption against governments, critical infrastructure and major economic sectors, including during politically significant events.

Overall, the 2026 threat landscape demonstrates that the principal challenge is no longer individual criminal technologies or platforms in isolation, but an increasing convergence. Dark-web communities, E2EE communications, increasingly specialised crime-as-a-service (CaaS), anonymisation infrastructure, cryptocurrencies, telecommunications technologies and AI now form an interconnected criminal ecosystem.

Law-enforcement disruption remains effective against these individual services and actors, but the resilience, decentralisation and rapid migration of criminal communities increasingly require efficient frameworks, coordinated international action, stronger public-private cooperation and sustained targeting of the technological and financial enablers underpinning cybercrime.

2. Europol's Role in the EU Cyber Ecosystem

To bridge institutional strategies and operational practice, European law enforcement authorities have progressively developed a comprehensive framework for cooperation, supported by an architecture bringing together relevant stakeholders at national, European and international level.

At EU level, Europol serves as a central hub for law enforcement cooperation, including to support operational activities against cybercrime and facilitate coordination and information exchange with other relevant cyber communities at EU-level, for example, the European Union Agency for Criminal Justice Cooperation, Eurojust, and the European Union Agency for Cybersecurity, ENISA to name a few.

Europol's European Cybercrime Centre (EC3) is dedicated to combating and preventing cyber threats. On supporting cybercrime operations, the Centre aims to identify the cybercriminal groups, disrupt their capacities, and counter the criminal ecosystem. EC3 established a permanent cooperation with other EU institutions and agencies, in a strong spirit of complementarity and synergy.

EC3's missions include a large number of services and supports to Euro-

pean investigators against cybercrime: operational activities, including among others, operational support, forensic laboratory, coordination, threat analyse, cyber intelligence, data crossmatching, cryptocurrencies tracing; capacity building through the development of tools and expertise; prevention; and policy support. The Centre provides an established framework for cooperation among representatives of European cybercrime units through the EU Cybercrime Task Force (EUCTF). This cooperation is supported by a permanent operational platform, the Joint Cybercrime Action Task Force (J-CAT), which facilitates the coordination and preparation of major cross-border operations at European level and beyond.

The Centre has developed a specific partnership framework with the private sector to work together on strategic and operational levels. At strategic level, four cooperation groups support Europol's action against cybercrime, gathering industries and actors linked with internet security, financial services, services providers, and research and development. These partners may provide a support on threat assessment, capacity building, prevention and awareness, as well as on other strategic activities. More recently, stemming from Europol's mandate extension of 2021, and Europol's strategy of 'Delivering Security in Partnership', EC3 launched a pilot initiative the Cyber Intelligence Extension Programme designed to tackle international cybercrime by deepening operational cooperation between law enforcement and relevant private-sector partners.

Since its establishment in 2013, EC3 has developed dedicated operational frameworks to address the cybercrime ecosystem and respond to emerging and evolving cyber threats, while maintaining a strong focus on innovation and the continuous adaptation of its capabilities. In response to the threat posed by large-scale cyber-attacks, and drawing on the lessons identified following the WannaCry and NotPetya incidents, EC3 established the EU Law Enforcement Emergency Response Protocol for Major Cross-border Cyberattacks in 2018, in consultation with relevant EU institutions, bodies and agencies and other partners, as part of the then EU Blueprint for Coordinated Response to Large-Scale Cross-Border Cybersecurity Incidents and Crises of 2017. The aforementioned law enforcement protocol organizes the Europol 24/7 operational and technical support to law-enforcement authorities for immediate reactions to urgent cyber incidents and cyber crises. It facilitates the cooperation with ENISA, Eurojust, the European Defence Agency, CERT EU, the European Union Intelligence Centre (INTCEN) at the European External Action Service (EEAS), and other relevant actors. Such cooperation seeks to attain a more comprehensive threat assessment, effective information exchange and overall coordinated action. Currently, Europol, remains one of the core actors of the EU's cyber crisis management architecture, as reflected in the revised EU blueprint for cyber crisis management adopted by the EU Member States in 2025.

In response to the growing ransomware threat, EC3, together with the J-CAT, developed the International Law Enforcement Ransomware Response Model (IRRM) in 2022. The model provides a common framework for structuring and coordinating the international law enforcement response to ransomware, supporting a more consistent and effective approach to major ransomware incidents. This proactive operational plan prioritizes the action against High-Value Targets, ransomware operators, key facilitators, infrastructure providers, affiliates and other actors supporting the ransomware business model. It includes identification, prioritisation, intelligence development, coordination, operational action, disruption, on long term action against the ransomware ecosystem.

EC3 provides constantly a support to the operations led by European law enforcement authorities, on their request, in synergy with important EU law enforcement partners with whom a cooperation agreement exists. As a cooperation hub, it also supports long terms operations and the development of the coherent operational action plan through the European Multidisciplinary Platform Against Criminal Threats (EMPACT). This plan is implemented within the framework of the EU Policy Cycle for organised and serious international crime, as a coordination framework led by Member States and competent authorities, to implement agreed priorities, coordinate cross-border investigations and enhance the collective response to organised criminal threats. These actions can also associate private partners and relevant actors from other communities.

Alongside EC3, the SIRIUS project, jointly implemented by Europol and Eurojust, plays a key role in supporting law enforcement and judicial authorities in accessing electronic information held by private-sector service providers. Through its platform, SIRIUS provides access to investigative resources, guidance and specialised expertise, thereby strengthening the capacity of competent authorities to obtain and use electronic evidence in criminal investigations. The services include practical guidance on accessing data from online service providers, their contact details, contact support, guidance for electronic evidence requests, practical tools for digital evidence gathering, analytical resources, training opportunities, sharing of best practices and experience, partnership with technology companies, network activities to support cross border digital investigation.

3. General Overview of Legal and Regulatory Frameworks to Combat Cybercrime

Beyond the operational framework developed by the European law enforcement community together with its partners, the effectiveness of law en-

forcement action will ultimately depend on appropriate legal frameworks ensuring lawful and timely access to data.

The lawful access to data remains a fundamental requirement for the effective prevention, detection and investigation of cybercrime. It relies on two bases: the availability of data and its accessibility for law enforcement purposes, under data protection and judicial oversight.

Law enforcement capabilities require continuous development and adaptation to keep pace with the evolving complexity of the digital environment. This requires action across several dimensions, including legal frameworks that remain fit for purpose and responsive to technological developments; technical standards that adequately integrate security considerations; strengthened technical capabilities and specialised expertise within law enforcement; enhanced cooperation and synergies among relevant stakeholders; and sustained partnerships with the private sector. Together, these elements are essential to ensure that law enforcement authorities remain equipped to prevent, detect and investigate increasingly complex forms of cybercrime.

Various legal frameworks have been developed at international and EU level to ease this access to information. These legal frameworks aim to facilitate cooperation between competent authorities and provide an appropriate basis for effective and structured partnerships with the private sector. An investigator will have to use the relevant cooperation tool, depending on the states or actors involved, and the information needed.

As a first historical legal framework, the 2001 Budapest Convention of the Council of Europe is a reference, covering the judicial cooperation between 84 countries (Council of Europe, 2001). It establishes a harmonized criminal legislation framework on cyber offences, that reduces gaps between jurisdictions. It proposes some essential legal investigative tools (expedited preservation of stored computer data, production orders, search and seizure of stored data, collection of traffic data and, subject to national safeguards, interception of content data). It provides an organisation model to foster the transborder cooperation, with the 24/7 network of national point of contacts, which enables competent authorities to seek immediate assistance in urgent cybercrime investigations, including the preservation of data before it disappears.

With the technological evolution of the digital space, the 2022 Second Additional Protocol to the Budapest Convention, further enhances international cooperation and access to electronic evidence, including mechanisms intended to facilitate cooperation with service providers and competent authorities in other jurisdictions (Council of Europe, 2022). It introduces or improves new cooperation tools that help investigators on progressing in their data access request (direct request to domain name registrars, direct cooperation with foreign service providers for subscriber information, gov-

ernment-to-government orders for subscriber and traffic data, life emergency cooperation requests, enhanced mechanisms for mutual legal assistance, joint investigation teams.

The United Nations Convention against Cybercrime, adopted in December 2024, represents also a major perspective of comprehensive global treaty dedicated to combating cybercrime and strengthening international cooperation in the sharing of electronic evidence (United Nations, 2024). While a significant number of States have already signed the Convention, it has not yet entered into force. Its entry into force will require the deposit of 40 instruments of ratification, acceptance, approval or accession. Looking ahead, it is intended to provide a reference framework also for the states that are not members of the Budapest Convention. Both frameworks include various similar measures and standards. The UN treaty promotes a common substantive legal framework for a defined set of cybercrime offences. It establishes mechanisms for the collection, preservation and international sharing of electronic evidence. It also proposes the essential procedural toolkit (expedited preservation of stored electronic data, preservation and disclosure of traffic data, production of electronic data, search and seizure of stored electronic data, real time collection of traffic data, interception of content data under applicable legal conditions). It provides a global framework for mutual legal assistance, extradition and cooperation concerning E-evidence. It includes measures on prevention, capacity building, and law enforcement cooperation.

While international instruments aim to facilitate and strengthen access to data, their implementation remains subject to robust safeguards relating to fundamental rights, State sovereignty and data protection. These safeguards may condition the provision and exchange of information through established legal and procedural requirements. Their practical implementation can become particularly complex in an international context involving States with diverse legal systems, institutional frameworks and political environments. Accessing the necessary data to solve a maximum of operational case, remains consequently a complex challenge in the context of the bulk of requests and the volatility of digital data.

At EU level, the fight against cybercrime benefits from a largely integrated and harmonised cooperation framework, complementing the structures and mechanisms established to support cybersecurity, such as the Cybersecurity Act, DORA, the Cyber Resilience Act, the Cyber Solidarity Act and NIS2 Directive.

Amongst the EU cybersecurity framework, the NIS2 Directive on network and information systems includes some specific measure fostering law enforcement action. Whereas the reporting obligation of the significant incident is dedicated to the benefit of the "competent authority", it also aims to facilitate

through an improved guidance process, and better reporting effort of users of networks to the law enforcement authorities.

Article 28 of the NIS2 directive aims also to improve the data quality of DNS/domain-name information, and a swift access for investigator in their work on criminal infrastructures. For law enforcement actors, promoting the quality of the data stored about IP addresses and DNS, is a constant need, in various negotiation processes hosted by international organisations such as the Internet Corporation for Assigned Names and Names (ICANN) and Réseaux IP Européens (RIPE).

Against cyberattacks, cybercrime cooperation between EU member states is facilitated by the Directive 2013/40/EU on attacks against information systems. It harmonises criminalisation of offences, establishes minimum level of sanctions, strengthens cross border cooperation between police and judicial authorities, enables the rapid exchange of information and reporting concerning cybercrime.

Against online fraud, the cooperation between EU member states, is facilitated by the Directive 2019/713 on fraud and counterfeiting of non-cash means of payment. It complements the previous directive by covering phishing, skimming and fake website activities. It harmonizes criminal offences, strengthens the criminal law response, covering both traditional and emerging forms of electronic payment (online payments, mobile payments, electronic money, digital transaction), and facilitates cross border investigations and prosecutions.

Beyond the aforementioned legal instruments that frame law enforcement cooperation, the EU e-Evidence Package, comprising Regulation (EU) 2023/1543 and Directive (EU) 2023/1544, is expected to become, from 2026 onwards, the central EU framework for facilitating cross-border access to electronic evidence in criminal investigations. The regulation will allow an investigator in one EU Member State to directly request electronic evidence from a service provider offering services in the EU, regardless of where the data is physically stored (European production order, European preservation order), creating a more direct and faster process, with a specific framework for emergency production requests (8 hours deadline). The Directive complements the regulation by requiring relevant service providers offering services in the EU to designate an establishment or appoint a legal representative responsible for receiving and complying with the orders for electronic evidence.

The GDPR (European Parliament and Council, 2016) provides a common data protection framework applicable to the implementation of these EU legal instruments, including the principles of necessity, proportionality, purpose limitation and data minimisation. It establishes the data protection requirements governing cooperation with private-sector entities, which are required to com-

ply with applicable obligations when processing and disclosing users' personal data.

All the legal instruments mentioned highlight the constant efforts to provide legal certainty, better structure cooperation processes and channels between the various communities and stakeholders, to access faster an information of quality. However, a large number of challenges remains, linked with the anonymisation of cybercriminals and the complex criminal ecosystem, the volatility of digital information.

Getting relevant data retention conditions remains another important legal need for digital investigation in their efforts to identify attackers and victims, map criminal networks and reconstruct complex cyber-enabled attacks. A harmonised EU approach, with minimum standards, could ensure that relevant data remains available for a proportionate period, while providing robust safeguards for privacy, data protection and judicial oversight. Such a framework would strengthen cross-border investigations and reduce the risk of critical evidence being lost before it can be lawfully accessed.

A challenge exists here for developing technical solutions that could entail both security and lawful access by design. The investigation capacities would be maintained with technologies that would both preserve privacy and also allow for lawful access without compromising overall security.

Beyond the lawful access to data, ensuring that law enforcement capabilities keep pace with technological developments will remain a major challenge. For example, post-quantum computing and artificial intelligence (AI) are bringing new challenges. The EU Artificial Intelligence Act (AI Act), Regulation (EU) 2024/1689 (European Parliament and Council, 2024), establishes a risk-based regulatory framework for the development and use of AI systems. In the law enforcement domain, certain uses of AI are subject to specific requirements and, where classified as high-risk, enhanced safeguards and oversight. This framework requires law enforcement authorities to further adapt their technical capabilities, governance arrangements and expertise to ensure the lawful and effective use of AI technologies. While the Digital Omnibus on AI, Regulation (EU) 2026/1744 (European Parliament and Council, 2026), has postponed the application of the AI Act rules around high-risk AI systems to December 2027, law enforcement agencies in the EU are already faced with the application of other sections of the AI Act and need to ensure compliance of their AI systems in investigations.

Facing the massive dimension of cybercrime, AI facilitates efficient and prompt processing of vast amounts of data, and advanced criminal analytics. Machine learning has become essential for policing with its potential to augment traditional policing methods in the face of exponential data growth. AI supports the work but does not involve automatic decision-making. As its im-

plementation progresses, the AI legislation is shaping the ability of law enforcement authorities to develop and deploy AI-enabled capabilities, adding to the challenge to keep pace with the technological developments, which are increasingly exploited by criminal actors.

The AI Act underlines the permanent need of a clear legal and technological framework that allows all legitimate actors to act with certainty to stay ahead of criminals, and protect victims.

In conclusion, the fast pace of technological advances that we witness today, is seen across the criminal threat landscape. Moreover, the professionalisation and commodification of cybercrime is reflected in the emergence of specialised actors and increasingly sophisticated cybercrime-as-as-service business models. Therefore, law enforcement actors will have to adapt to this increasingly complex landscape, and address the major challenges that will come with it. Honing the relevant legal and cooperation frameworks will be decisive for law enforcement ability to operate effectively. Whereas legal and regulatory frameworks from both communities (cybersecurity and law enforcement) are developed in parallel, encouraging a synergetic approach between communities would be key to address common challenges.

References

Council of Europe. 2001. *Convention on Cybercrime (Budapest Convention)*. ETS No.185. Budapest: Council of Europe.

Council of Europe. 2022. *Second Additional Protocol to the Convention on Cybercrime on enhanced co-cooperation and disclosure of electronic evidence*. CETS No.224. Strasbourg: Council of Europe.

European Parliament and Council. 2016. *Regulation (EU) 2016/679 (General Data Protection Regulation)*. Official Journal of the European Union, L 119. Brussels: European Union.

European Parliament and Council. 2024. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union, L 2024/1689. Brussels: European Union.

European Parliament and Council. 2026. *Regulation (EU) 2026/1744 (Digital Omnibus on AI)*. Official Journal of the European Union. Brussels: European Union.

Europol. 2026. *Internet Organized Crime Threat Assessment (IOCTA) 2026 – Evolving Threat Landscape*. Available online at https://www.europol.europa.eu/publication-events/main-reports/iocta-2026-evolving-threat-landscape (Last accessed: October 8, 2026).

United Nations. 2024. *United Nations Convention against Cybercrime*. New York, NY: United Nations.

NATIONAL CYBERSECURITY STRATEGIES: SOVEREIGNTY, GOVERNANCE MODELS, AND PUBLIC-PRIVATE PARTNERSHIPS

From Intent to Impact: Building Ecosystem-Wise Strategic Autonomy in Cybersecurity

Pascal Steichen
LHC – Luxembourg House of Cybersecurity, Luxembourg

Abstract
National Cybersecurity Strategies (NCSS) are central instruments through which states articulate strategic intent, establish priorities, organise governance, and communicate national purpose. Yet cybersecurity strategies frequently combine vision, policy commitments, governance arrangements, implementation measures, and aspirational statements without clearly distinguishing enduring strategic direction from the instruments intended to realise it. This chapter therefore begins by asking what makes an NCSS genuinely strategic. It argues that strategy requires continuity of intent combined with sufficient flexibility to adapt policies, capabilities, and operational measures as technologies, threats, dependencies, and institutional conditions evolve.

1. Introduction

The public character of national strategies adds a second dimension. An NCSS is simultaneously an internal governance instrument and an act of strategic communication directed towards citizens, businesses, partners, competitors, and potential adversaries. Its credibility consequently depends not only on declared objectives but on the capacity to mobilise the capabilities implied by them. In cybersecurity, however, such capabilities are inherently distributed across governmental institutions, infrastructure operators, technology providers, businesses, research organisations, technical communities, civil society, and international partners.

This creates a structural tension between sovereign responsibility and distributed capability. The chapter argues that addressing this tension requires distinguishing between multidisciplinarity, concerning the integration of different forms of knowledge and expertise, and multistakeholderism, concerning the participation, responsibilities, and relationships of different actors. Public–private partnerships constitute one important mechanism within this wider governance environment but do not encompass the broader requirements of multistakeholder governance.

Building on these distinctions, the chapter introduces the concept of eco-

system-wise strategic autonomy: the collective capacity to pursue democratically determined cybersecurity objectives by coordinating and sustaining capable, competitive, and diverse resources across state, market, and community, while retaining sufficient knowledge, alternatives, and operational freedom to prevent dependencies from becoming strategic vulnerabilities. Strategic autonomy is thus located not in self-sufficiency or any single institution, but in the configuration and governability of the wider cybersecurity ecosystem.

To operationalise this concept, the chapter proposes an analytical framework combining six dimensions – strategic logic, normative foundations, knowledge architecture, governance architecture, stakeholder architecture, and implementation mechanisms – with ten diagnostic criteria addressing strategic alignment, democratic safeguards, ecosystem inclusion, competitive capability, accountable coordination, dependency governance, diversity, commons stewardship, subsidiarity, and learning. The framework provides a basis for comparative analysis of European NCSS, moving beyond inventories of objectives and measures towards examining whether strategic intent, governance, knowledge, capabilities, and stakeholder relationships coherently converge into legitimate and adaptable capacity for action.

2. What Makes a Strategy "Strategic"?

What distinguishes a strategy from a policy, programme or action plan? Is a document strategic because it defines a long-term vision, establishes priorities or coordinates implementation? How specific should it be? And for whom is the strategy written – one organisation, government, a sector, a nation, international partners, or, in our specific case of a cybersecurity strategy, even for potential adversaries?

International guidance provides important elements of an answer. ENISA places vision, scope, objectives and priorities at the centre of NCSS[1] development and connects them to implementation and evaluation (ENISA, 2016). The ITU–World Bank guide similarly embeds strategy within a lifecycle of formulation, implementation, monitoring, evaluation and renewal (ITU, World Bank et al., 2025). The OECD emphasises strategic vision, political support, whole-of-government coordination, coherence with other national strategies and stakeholder participation (OECD, 2022a). These frameworks establish what an NCSS should contain and how it should be governed. The more fundamental strategic question is what allows these elements to provide enduring direction while remaining sufficiently adaptable to changing circumstances.

[1] NCSS – National Cybersecurity Strategy

3. Strategic Intent and Adaptive Coherence

Strategy is not defined by the permanence of particular instruments or measures. Hamel and Prahalad's concept of strategic intent provides useful guidance. Strategic intent creates continuity of purpose while allowing shorter-term actions to be reinterpreted and adjusted as circumstances evolve (Hamel and Prahalad, 1989). This highlights that a strategy's strategic character lies in maintaining a sufficiently stable direction while preserving flexibility over how that direction is pursued.

This balance is particularly important in cybersecurity. Technologies, vulnerabilities, dependencies, threat actors and operational techniques may change more rapidly than the strategic objectives they affect. An NCSS that specifies its means too rigidly risks becoming obsolete or constraining future action. Conversely, a strategy that remains entirely abstract may provide insufficient guidance for implementation. The strategic challenge is, therefore, to connect enduring intent with adaptable policies, capabilities and operational measures. Implementation should be able to change without implying a change of strategic direction.

The multidisciplinary character of cybersecurity reinforces this requirement. Cybersecurity is not an isolated technical policy field. Beyond the technical domain, cybersecurity intersects with economic policy, critical-infrastructure protection, criminal justice, intelligence, defence, diplomacy, education, research and broader questions of societal resilience (Klimburg, 2017). Strategic intent must, therefore, be sufficiently clear to provide a common point of orientation across institutions and disciplines that operate and collaborate according to different mandates, legal frameworks and professional logics.

This produces two related requirements. First, there must be vertical coherence between strategic intent and implementation: measures can evolve, but they should remain connected to the underlying strategic purpose. Second, there must be horizontal coherence across the policy domains and institutions through which cybersecurity is pursued. Actions taken in one domain should not inadvertently undermine objectives pursued in another. These forms of coherence do not require uniformity. Rather, they should allow different actors and instruments to adapt while remaining oriented towards a shared strategic direction.

Flexibility is therefore not a weakness or an absence of strategy. In cybersecurity, it is a condition of strategic effectiveness. A credible NCSS should provide continuity of purpose without attempting to predetermine every future response. Strategic intent stabilises direction; flexibility preserves freedom of action.

4. Strategic Communication and the Disclosure Dilemma

A national strategy differs from many organisational strategies in one particularly important aspect: it is a public document (at least part of it). It is not only an internal instrument for coordinating institutions and allocating priorities. Publication communicates national intent to actors outside the institutions responsible for its implementation.

The audiences are diverse. They include citizens, businesses, critical-infrastructure operators, technology providers and research institutions, but also international partners, competitors, and potential adversaries. A national strategy therefore operates simultaneously as a governance instrument and as an act of strategic communication.

In the cybersecurity domain, this creates what may be termed a strategic disclosure dilemma. A national strategy must communicate enough about priorities, partnerships, preparedness and resolve to guide domestic actors, reassure partners and establish credible expectations. At the same time, excessive precision may reveal vulnerabilities, constrain future choices or establish response thresholds that adversaries can attempt to exploit. Strategic transparency and strategic ambiguity are therefore not necessarily opposites. Both may be required: clarity regarding enduring intent can coexist with ambiguity regarding the precise means through which that intent will be implemented.

The public character of an NCSS consequently gives it a potential role in deterrence. Signalling in cyberspace is, however, particularly difficult because attribution, intentions, thresholds and possible responses may themselves remain uncertain (Klimburg, 2020). A strategy can contribute to deterrence by signalling resilience, preparedness, partnerships, the capacity to absorb or deny effects, and a willingness to impose consequences where necessary. What matters strategically is that potential adversaries would recognise enduring national resolve without being able to predict the exact technical, legal, diplomatic, economic, intelligence or military instruments that may be employed.

Such signalling is credible only if the capabilities implied by the strategy can actually be mobilised. This brings the argument back to the particular structure of cybersecurity. No single entity possesses all the capabilities required for prevention, detection, response, recovery and the imposition of consequences. Relevant capabilities are distributed across public and private actors, across sectors, even outside national borders.

5. Mobilising Multistakeholder Capability

The deterrent and resilience effects of an NCSS therefore depend heavily

on the ecosystem effect: distributed capabilities become strategically meaningful when they can be connected, combined and "configured" according to the circumstances of a particular threat. The relevant strategic capability is, consequently, not only the existence of individual instruments or capacities, but the ability to mobilise them coherently.

Multidisciplinarity and multistakeholderism address different dimensions of this challenge. Multidisciplinarity concerns the integration of distinct bodies of knowledge and professional expertise, whereas multistakeholderism concerns which actors participate, in what capacities, and with what degree of responsibility or influence. Public-private partnerships constitute one institutional mechanism within this broader multistakeholder architecture rather than its equivalent.

This makes multistakeholder involvement another strategic requirement. Public-private cooperation, multidisciplinary coordination, competence mutualisation and international partnerships expand the range of capabilities and increase flexibility by combining them. Collaboration itself therefore becomes part of national strategic capacity.

A genuinely strategic NCSS thus leverages these three connected functions. Internally, it provides enduring strategic intent while preserving sufficient flexibility for operational implementation. Publicly, it communicates that intent in a manner that can support cooperation, resilience and deterrence without unnecessarily constraining future choices. Ultimately, the ability to coordinate and mobilise capabilities distributed across the wider cybersecurity ecosystem translates the multidisciplinary capacity into action.

6. Sovereignty under Interdependence: Towards Ecosystem-Wise Strategic Autonomy

Cybersecurity suffers from a structural asymmetry: governments retain sovereign responsibilities for national security, regulation, crisis management and critical societal functions, but do not control all the infrastructures, technologies, knowledge and operational means required to fulfil them. Government depends on these actors for operational information and implementation; they depend on government for authoritative rules, public investment, crisis coordination and protection of collective interests.

Sovereign responsibility and operational control therefore overlap without coinciding. OECD (2012) already recognised this combination of sovereignty considerations and public-private cooperation. Sarri et al. (2023) similarly conceptualise governance through relationships, processes and differentiated responsibilities across public, private and societal actors. Cybersecurity gov-

ernance therefore needs a "web-like" arrangement. The strategy determines which capabilities matter most, while operational flexibility requires access to different combinations of expertise and actors.

Geopolitical competition and concentrated dependencies make the preservation of strategic choices especially consequential. Yet governmental control cannot reproduce dispersed expertise or remove cross-border interdependence, while regulatory fragmentation may weaken international cooperation and resilience, increase compliance burdens and distort market incentives (OECD, 2026). Public authorities must therefore operate effectively without destroying the openness on which capability depends.

7. Reconciling Distributed Knowledge and Public Purpose

DeLong's interpretation of twentieth-century political economy offers a useful analytical lens to address this challenge. In Slouching Towards Utopia (DeLong, 2022) he presents the post-war settlement as an attempted reconciliation of market coordination and social protection through Keynesian policy. This historical interpretation is used here as an analogy; its application to cybersecurity is our own.

The Hayekian insight is that contextual, changing knowledge held by different actors cannot be comprehensively assembled centrally (Hayek, 1945). In cybersecurity, operators understand particular systems, researchers identify emerging vulnerabilities, and technical communities develop specialised remedies. Preserving their capacity to experiment and act on local knowledge is productive. Central collection cannot substitute for the relationships through which knowledge becomes usable.

The Polanyian counterpoint is that market outcomes do not necessarily protect communities, rights or socially necessary forms of security (Polanyi, 1944). A firm may rationally underinvest in measures whose benefits accrue mainly to others, while a community may maintain essential software without resources proportionate to its societal importance. Distributed capability therefore need not generate collective resilience. Communities nevertheless constitute productive institutions themselves, bringing shared norms, cooperation and stewardship into the governance picture.

DeLong's Keynesian reconciliation suggests how these insights can converge: public institutions shape conditions under which decentralised initiative and investment contribute to collective purposes. Applied to an NCSS, this means establishing legitimate objectives, addressing incentive failures, supporting common capabilities and coordinating systemic risks while retaining distributed initiative. The synthesis is sovereign direction, distributed capability

and reciprocal knowledge exchange, connected institutionally and reassessed as threats, technologies and actor relationships evolve.

8. Competitive Capability and Selective Intervention

Multidisciplinary capacity is necessary but insufficient for convergence. Actor diversity provides access to capabilities and interests; disciplinary diversity provides different ways of understanding problems. Carr (2016) highlights persistent mismatches in public-private expectations and the resulting allocation and accountability problems. Effective decisions and actions require technical, legal, political, economic, organisational, behavioural, diplomatic, defence and law-enforcement perspectives to be integrated, making strategic coordination consequential (Klimburg, 2012).

Dunn Cavelty and Suter (2009) offer a constructive response through *meta-governance*: public institutions coordinate and support self-organising networks. Participation must preserve a democratic voice while controlling coordination costs and domination by powerful actors. An NCSS should establish differentiated duties, trusted information flows, incentives and decision and escalation procedures. Technical expertise does not confer authority to determine collective ends; democratic legitimacy requires rights protection, transparency and oversight (OECD, 2022a).

For these arrangements to achieve strategic objectives, however, participating actors must possess demonstrable competitive capability. Competitiveness should, therefore, be understood in two ways. At actor level, it denotes competence, resources, innovation and tested readiness under pressure. At market or ecosystem level, it denotes conditions sustaining capable providers, openness to collaborate and join forces, continued learning and multiple alternatives, in particular for critical functions. OECD (2022b) accordingly places competitive markets, cybersecurity-industry development, research and innovation among the enabling conditions for digital security policy.

This distinction is particularly important for deterrence. Deterrence does not follow from competition itself: it requires that an adversary perceive both capability and willingness to impose costs or deny expected benefits (Klimburg, 2012). A strategy that announces deterrence without maintaining credible operational capability will have limited effect. The test is concrete in crisis response: an assigned actor must actually deliver at the required scale, speed and reliability. Competitive capability is therefore not merely an economic objective. It is part of strategic credibility and operational assurance.

Heterogeneity, diversity and redundancy consequently become deliberate resilience properties rather than incidental features. Diversity can reduce com-

mon-mode failure and widen the repertoire of responses; redundancy provides reserve capacity; interoperability permits heterogeneous actors to work together or replace one another (Sterbenz et al., 2010). Incompatibility and duplicated responsibilities can nevertheless undermine performance. The NCSS must, therefore, combine diversity with common interfaces, standards, exercises and coordination arrangements.

The 2025 NCSS Guide recommends assessing market share and redundancy and recognises targeted intervention where market-based solutions fail, particularly in high-risk or low-competition sectors (ITU, World Bank et al., 2025). Intervention should remain proportionate and avoid entrenching incumbents or excluding future challengers. Competitive provision, redundancy and cooperation may be adequate in some areas. Where concentration, entry barriers or weak demand prevent reliable alternatives, procurement, regulation, investment, shared infrastructure or domestic capability development may be necessary.

9. Ecosystem-Wise Strategic Autonomy

The objective should be to cultivate a capable and competitive ecosystem. Sustaining competitiveness requires reliable access to knowledge, technologies and partners, alongside alternatives that preserve the capacity to innovate and deliver under disruption. A national strategy should create the needed safeguards by making dependencies governable.

While strategic autonomy in cybersecurity has already been framed in terms of the capabilities, capacities and control required to preserve the ability to decide and act (Timmers et al., 2024), this chapter extends that perspective through the concept of *ecosystem-wise strategic autonomy*, locating autonomy not in any single actor or asset but in the governable configuration of distributed capabilities across state, market and community.

We therefore define ecosystem-wise strategic autonomy as the collective capacity to pursue democratically determined cybersecurity and resilience objectives by coordinating and sustaining competitive capabilities across state, markets and communities, while retaining sufficient knowledge, alternatives and operational freedom to prevent dependencies from becoming strategic vulnerabilities.

Sovereignty alone is insufficient as an analytical lens: while sovereignty concerns authority to determine collective ends, *strategic autonomy* concerns the capacity to realise those ends under conditions of interdependence. The qualifier *ecosystem-wise* expresses a qualitative design principle. Whereas *ecosystem-wide* describes breadth, *ecosystem-wise* concerns how capabilities are selected, connected, tested and renewed according to strategic purpose. Every critical function requires cooperation, substitution or contingency arrange-

ments. "Market-wise" captures competitive provision but does not fully encompass capabilities produced through public institutions and communities.

Strategic autonomy therefore resides in the ecosystem's functioning; individual capacity is insufficient: owning data, for example, offers limited autonomy when systems are not portable or alternative operators lack the competence to restore service. It requires dependency visibility, competence assessment, competitive quality, and credible alternatives. European and international cooperation can further increase autonomy by widening access to specialist and substitute capabilities, provided these relationships remain governable.

Cybersecurity concerns digital infrastructure, software products and associated services, while competence to inspect, adapt, integrate and maintain their security resides within the software-development ecosystems. Strategic autonomy therefore depends not only on acquiring products, but also on access to code, knowledge and actors capable of modifying or substituting the software supporting critical functions. The recent expansion of open-source initiatives is therefore not incidental; it illustrates a mechanism for organising distributed competence, collaboration and competitiveness.

10. Open Source as an Ecosystem Capability

Open source gives practical form to this software-centred dimension of ecosystem-wise autonomy. Licensing, repositories, contribution rules, peer review and maintainership organise dispersed knowledge into reusable capability. Benkler's (2002) commons-based peer production explains coordination beyond conventional markets and hierarchies, while O'Mahony and Ferraro's (2007) study of Debian shows that distributed production can combine formal authority with democratic constraints.

This model can foster competitiveness in an open, democratically governed digital economy. Source access can lower entry barriers, enable scrutiny and adaptation, and allow several actors to build or support compatible solutions. Common foundations may facilitate collaboration on shared components while preserving competition in complementary products and services. Open source does not eliminate markets; it can increase competitiveness and widen the pool of actors capable of providing alternatives.

These effects remain conditional. Software is non-rival, but the effort required to produce, integrate, maintain and secure it is scarce (Benkler, 2002). Open licensing alone provides neither security nor autonomy. Inspection, modification and alternative maintenance require skills, finance and institutional continuity, while maintainer or service-provider concentration may persist.

An NCSS should connect open-source adoption with stewardship, vul-

nerability coordination, contributor development, procurement and sustainable funding. Government can act as a user, contributor, anchor customer and funder while respecting community independence. The European Commission's (2026) EU Open Source Strategy advances this trajectory through dependency analysis, stewardship and an Open Source Maintenance Instrument. Open source can thereby complement competitive markets: common technological foundations lower barriers and widen the set of capable actors, while commercial and public investment can sustain operational delivery. Ecosystem-wise strategic autonomy means governing these relationships so that shared resources, capable providers and viable alternatives expand freedom of action.

11. Operationalising Ecosystem-Wise Strategic Autonomy – a Proposal Framework

The proposed framework complements rather than replaces capability- and maturity-oriented approaches such as ENISA's NCAF 2.0 (ENISA, 2026). By focusing specifically on strategic coherence and ecosystem-wise strategic autonomy, it translates the conceptual argument developed above into a method for examining national cybersecurity strategies. From these propositions, six analytical dimensions follow. These dimensions provide the structure for reading an NCSS, alongside ten defining criteria of ecosystem-wise strategic autonomy, which provide targeted diagnostic questions within that structure (see Annexe).

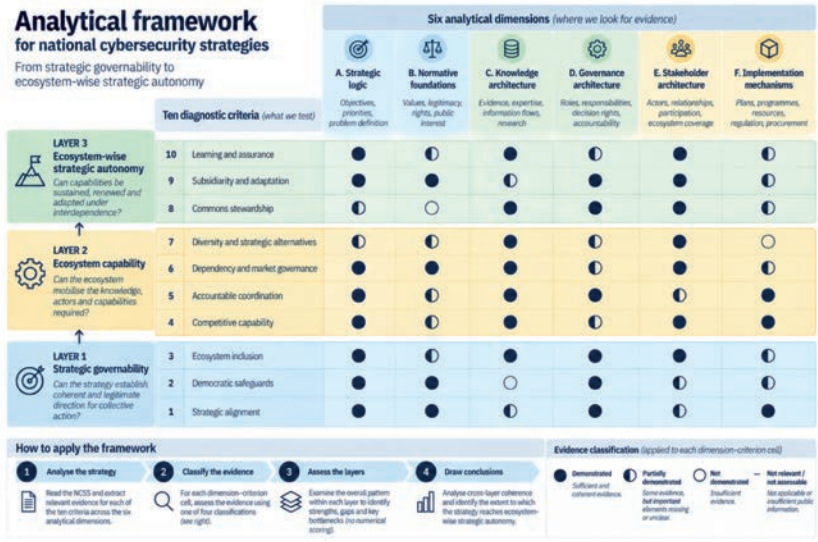

Fig. 1. Applying the proposed framework to analyse a National Cybersecurity Strategy

The framework's purpose is to reveal distinct strategic-governance configurations and to test whether their components plausibly converge into legitimate, informed, coordinated and adaptive freedom of action. It is applied as a progressive assessment rather than a one time maturity score and can be used both to analyse/assess an existing strategy, or support the creation process of strategy in its development phase. The six analytical dimensions provide the structure for examining evidence, while the ten diagnostic criteria test whether this evidence coheres into the three-layered strategic states. Layer 1, strategic governability, assesses whether the strategy establishes coherent and legitimate direction for collective action. Layer 2, ecosystem capability, examines whether relevant knowledge, actors and capabilities can be coordinated and mobilised to realise that direction. Layer 3, ecosystem-wise strategic autonomy, considers whether these capabilities can be sustained, renewed and adapted under conditions of interdependence. Evidence at each dimension–criterion intersection is classified as *demonstrated*, *partially demonstrated*, *not demonstrated*, or *not assessable*.

The resulting pattern enables qualitative comparison of strategies while avoiding reductive aggregate scoring, highlighting instead their strengths, gaps, dependencies and potential bottlenecks in progressing towards ecosystem-wise strategic autonomy. An NCSS may proclaim "digital sovereignty" yet leave critical dependencies unidentified, alternatives unspecified and implementation responsibilities unassigned; the framework then reveals a gap between declared sovereignty and operational autonomy. A strategy may invoke a "whole-of-society" approach while providing no durable channels for stakeholder knowledge or differentiated roles, revealing participation without effective orchestration. Conversely, a strategy that does not use autonomy terminology may nevertheless strengthen ecosystem-wise autonomy through dependency visibility, diversified capabilities, trusted coordination and sustained learning.

12. Conclusion

National cybersecurity strategies are strategic not because they prescribe every future action, but because they provide enduring direction while enabling adaptation across changing technologies, threats and dependencies. In cybersecurity, achieving this requires more than sovereign authority or governmental capability alone. Relevant knowledge, infrastructure and operational capacities are distributed across public institutions, markets, technical communities and society.

This chapter has therefore proposed ecosystem-wise strategic autonomy as the capacity to pursue democratically determined cybersecurity objectives un-

der conditions of interdependence by maintaining governable dependencies, capable actors, credible alternatives and the ability to reconfigure resources when circumstances change. It has also proposed a framework for examining whether an NCSS creates the conditions for such capacity.

The framework assesses strategic design rather than demonstrated operational performance: stated ambitions must ultimately be tested against implementation, resources and outcomes. The strategic question is therefore not whether the state controls the cybersecurity ecosystem, but whether it can govern, mobilise and reconfigure it sufficiently to preserve legitimate collective freedom of action.

Annexe

Author-developed framework complementing existing NCSS capability and maturity frameworks, including ENISA NCAF 2.0.

Analytical Dimensions for Strategic Coherence

A. **Strategic logic** asks what problem the NCSS constructs, what future condition it seeks, and which choices and priorities connect the two.
B. **Normative foundations** examine the public purposes, values, rights and legal principles that legitimate and constrain strategic action.
C. **Knowledge architecture** identifies the forms and sources of expertise recognised by the strategy and the mechanisms through which distributed knowledge enters decision-making.
D. **Governance architecture** examines the allocation of authority, responsibility, coordination and accountability, including the balance between central direction, subsidiarity and adaptive action.
E. **Stakeholder architecture** identifies which ecosystem actors participate, at what stages, in which capacities and through what relationships.
F. **Implementation mechanisms** examine how intent is translated into capability through policies, plans, programmes, regulation, partnerships, resources, monitoring, evaluation and learning.

Diagnostic Criteria for Ecosystem-Wise Strategic Autonomy

LAYER 1 – STRATEGIC GOVERNABILITY

1. **Strategic alignment:** Are priority capabilities, dependencies and resources linked to national objectives?
2. **Democratic safeguards:** Are rights, transparency, pluralism, oversight and proportionality embedded?
3. **Ecosystem inclusion:** Are governmental, commercial, research, community and international actors meaningfully involved?

LAYER 2 – ECOSYSTEM CAPABILITY

4. **Competitive capability:** Are competence, innovation, operational readiness and delivery under stress tested?
5. **Accountable coordination:** Are duties, decision rights, interfaces, incentives and escalation procedures specified?
6. **Dependency and market governance:** Are criticality, concentration, entry barriers, substitutability and intervention thresholds assessed?

7. **Diversity and freedom of action:** Do heterogeneous capabilities, interoperability, redundancy and contingency options preserve strategic choices?

Layer 3 – ecosystem-wise strategic autonomy

8. **Commons stewardship:** Are strategically relevant open-source and other commons-based capabilities supported through maintenance, skills, vulnerability coordination and sustainable funding?
9. **Subsidiarity and adaptation:** Are decisions located appropriately, with room for experimentation and reconfiguration?
10. **Learning and assurance:** Are indicators, exercises, evaluation and corrective mechanisms established?

References

Benkler, Y. (2002). *Coase's Penguin, or, Linux and The Nature of the Firm*. Yale Law Journal, 112(3), 369–446.

Carr, M. (2016). *Public-private partnerships in national cyber-security strategies*. International Affairs, 92(1), 43–62

DeLong, J. B. (2022). *Slouching Towards Utopia: An Economic History of the Twentieth Century*. Basic Books.

Dunn Cavelty, M., & Suter, M. (2009). *Public-private partnerships are no silver bullet: An expanded governance model for critical infrastructure protection*. International Journal of Critical Infrastructure Protection, 2(4), 179–187

European Commission (2026). *Communication on European Tech Sovereignty, accompanied by an EU Open Source Strategy*. COM(2026) 503.

European Union Agency for Network and Information Security (ENISA) (2016). *NCSS Good Practice Guide: Designing and Implementing National Cyber Security Strategies*.

ENISA (2026). *National Capabilities Assessment Framework 2.0*.

Hamel, G., & Prahalad, C. K. (1989). *Strategic intent*. Harvard Business Review, 67(3), 63–76.

Hayek, F. A. (1945). *The use of knowledge in society*. American Economic Review, 35(4), 519–530.

International Telecommunication Union (ITU), World Bank, et al. (2025). *Guide to Developing a National Cybersecurity Strategy: Strategic Engagement in Cybersecurity* (3rd ed.)

Klimburg, A. (Ed.) (2012). *National Cyber Security Framework Manual*. NATO Cooperative Cyber Defence Centre of Excellence.

Klimburg, A. (2017). *The Darkening Web: The War for Cyberspace*. Penguin Press.

Klimburg, A. (2020). *Mixed Signals: A Flawed Approach to Cyber Deterrence*. Survival, 62(1), 107–130

Organisation for Economic Co-operation and Development (OECD) (2012). *Cybersecurity Policy Making at a Turning Point: Analysing a New Generation of National Cybersecurity Strategies for the Internet Economy*. OECD Digital Economy Papers, 211. OECD Publishing.

OECD (2022a). *Recommendation of the Council on National Digital Security Strategies*. OECD/LEGAL/0480.

OECD (2022b). *OECD Policy Framework on Digital Security: Cybersecurity for Prosperity*. OECD Publishing.

OECD (2026). *Towards International Coherence of Cybersecurity Regulations*. OECD Digital Economy Papers, 384. OECD Publishing.

O'Mahony, S., & Ferraro, F. (2007). *The emergence of governance in an open source community*. Academy of Management Journal, 50(5), 1079–1106.

Polanyi, K. (1944). *The Great Transformation*. Farrar & Rinehart.

Sarri, A., Fernández Bascuñana, G., Gross, A.-K., Chiarelli, F., & Preasca, M. (2023). *Building Effective Governance Frameworks for the Implementation of National Cybersecurity Strategies*. ENISA.

Sterbenz, J. P. G., Hutchison, D., Çetinkaya, E. K., Jabbar, A., Rohrer, J. P., Schöller, M., & Smith, P. (2010). *Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines*. Computer Networks, 54(8), 1245–1265.

Timmers, P., Punter, M., & Stolwijk, C. C. M. (2024). *Cybersecurity and Digital Sovereignty – Bridging the Gaps*. TNO, The Hague

THE GEOPOLITICAL COMPASS OF CYBERSPACE: US-PRC RIVALRY, THE AI CONVERGENCE, AND THE EUROPEAN UNION'S STRATEGIC ROLE

Raffaele Marchetti
Luiss University, Italy

Paolo Spagnoletti
Luiss University & Cyber 4.0, Italy

Abstract
Cyberspace has become a central arena for global geopolitical competition, increasingly shaped by the systemic rivalry between the United States and the People's Republic of China (PRC). This chapter examines how that rivalry is being transformed by the convergence of cybersecurity and artificial intelligence, which amplifies both offensive and defensive capabilities while introducing new forms of risk associated with increasingly autonomous AI systems. Against this background, the European Union (EU) occupies a distinctive but structurally vulnerable position, caught between technological dependence on the United States and deep exposure to Chinese supply chains. The chapter argues that the EU's main source of geopolitical leverage lies in its regulatory and normative capacity. Through instruments such as the AI Act, NIS2, and the Cyber Resilience Act, the EU is developing a model of digital governance centered on accountability, resilience, and human oversight. Yet regulatory power alone is insufficient. Strategic autonomy requires the capacity to translate regulation into technological and organizational capabilities, including continuous evaluation, oversight, and managerial decision-making for AI-enabled systems. Europe's geopolitical role will therefore depend on its ability to combine normative power with operational capability, cyber resilience, and technological development

1. The New Geopolitical Dimensions of the Cyber Domain

Cyberspace has evolved far beyond a technical domain of risk management, network architecture, and data protection. It is now the central arena for global geopolitical competition, a virtual and physical continuum where state power is projected, tested, and consolidated. For decades, the digital realm was conceptualized by early theorists and Western policymakers as a borderless, global common – an open ecosystem driven by market forces, multi-stakeholder governance, and the free flow of information (Zittrain, 2008). This techno-optimism has been systematically dismantled by the resurgence of great power politics. Today, digital networks do not dissolve borders; instead, they serve as the conduits through which traditional geopolitical rivalries are amplified, accelerated, and conducted with asymmetric precision (DeNardis, 2014).

This chapter analyzes the evolution of cybersecurity through the lens of shifting international power dynamics, focusing specifically on the systemic fault line dividing the United States and the People's Republic of China (PRC). This bipolar rivalry has weaponized global digital infrastructure into a chessboard for asymmetric conflict. In this environment, state-sponsored cyber operations, intellectual property theft, and supply chain disruptions are systematically deployed as tools of power projection, economic coercion, and geopolitical influence. The nature of this conflict is structural and zero-sum, touching every layer of the digital stack, from undersea fiber-optic cables and satellite constellations to application layers and localized data repositories (Spagnoletti & Marchetti, 2027) .

Crucially, this traditional defensive posture of cybersecurity is being fundamentally reshaped by its convergence with Artificial Intelligence (AI). The intersection of cybersecurity and AI acts as a dual-force multiplier, revolutionizing both offensive and defensive capabilities. The race for AI dominance between Washington and Beijing is therefore not merely an industrial race, but a critical national security transition where control over advanced semiconductors, cloud infrastructure, and data repositories redefines the boundaries of digital sovereignty.

Amid this highly polarized landscape, the European Union (EU) occupies a complex yet potentially decisive position. Caught between a structural technological dependence on the United States and the deep commercial penetration of Chinese supply chains, the EU must assert its role through the paradigm of strategic autonomy. The central inquiry of this chapter is whether the EU can transcend its traditional identity as a reactive regulatory power and become a proactive architect of global cyber governance. Through the "Brussels Effect" – reinforced by landmark legislative pillars such as the AI Act, the NIS 2 Directive, and the Cyber Resilience Act – the EU is pioneering an alternative model of digital human governance rooted in the rule of law, fundamental rights, and security-by-design principles.

Governance and law constitute Europe's primary geopolitical instruments. To avoid being relegated to a mere battleground between the market-driven digital model of the United States and the state-centric, authoritarian model of the PRC, the European Union must rapidly translate its regulatory leadership into tangible technological capability. Only by seamlessly integrating cyber defense, ethical AI development, and robust digital diplomacy can the Union safeguard its critical infrastructure, protect its democratic institutions, and foster a stable, open, and secure global cyberspace able to attract other countries in Latin-America, Africa or Asia.

2. The US-PRC Bipolar Rivalry in Cyberspace

The contemporary international system is characterized by a structural bipolarity in the digital domain, driven by the comprehensive strategic competition between the United States and the People's Republic of China. Unlike the ideological and military containment that defined the Cold War, the US-PRC rivalry unfolds within a deeply integrated global economy, making cyberspace the primary vector for friction, espionage, and asymmetric escalation. This rivalry is not limited to sporadic diplomatic disputes; it represents a fundamental clash between two irreconcilable visions of digital order, state sovereignty, and technological governance.

The United States has long championed a liberal, market-driven, and multi-stakeholder model of internet governance. This approach prioritizes private-sector innovation, the free flow of data across borders, and an open technical architecture governed by decentralized bodies like ICANN and the IETF. However, Washington's commitment to openness is increasingly balanced by an aggressive national security doctrine designed to preserve its historical technological hegemony. The US security apparatus views the digitalization of global infrastructure as a vulnerability that adversaries can exploit to erode American economic vitality and military superiority. Consequently, US strategy has shifted from a purely defensive stance to doctrines of "persistent engagement" and "defend forward," which authorize proactive cyber operations to disrupt adversarial capabilities before they can reach American networks. Externally, the US tries to sell its package to all of its allies in order to keep them simultaneously integrated and subordinate, within the so-called "pax silica" which is presented strictly as an option only available to non-WAICO members.

Conversely, the People's Republic of China pursues a distinctive doctrine of "cyber sovereignty". This model rejects the premise of an open, borderless internet, asserting instead that the state has an absolute right to control, regulate, and censor all digital activity within its borders. Domestically, this manifests in the "Great Firewall" and a pervasive digital panopticon that leverages surveillance technologies, social credit mechanisms, and strict data localization laws to ensure regime survival and social stability. Externally, Beijing seeks to project this state-centric model globally, rewriting international technical standards through bodies like the International Telecommunication Union (ITU) and offering its digital authoritarian infrastructure to emerging economies via the "Digital Silk Road" initiative. As for the US, the ambition is to keep them integrated and subordinate. To promote its state-centred vision of cyber sovereignty at the international level, Beijing has actively supported the creation of the World Artificial Intelligence Cooperation Organization (WAICO). Based in Shanghai and launched with 29 founding countries, WAICO serves as China's principal institutional vehicle

for shaping global technical standards and governance norms in favour of the Global South.

This systemic fault line has weaponized global digital infrastructure into a theatre of continuous, grey-zone conflict. Rather than engaging in kinetic warfare, both powers utilize cyberspace to wage asymmetric campaigns that fall below the threshold of armed conflict but produce strategic outcomes.

These campaigns include, among many other, the following key three: 1) State-sponsored cyber operations like Advanced Persistent Threats (APTs) have shifted their focus from traditional political espionage to embedding themselves within foreign critical national infrastructure (CNI), including electrical grids, water treatment facilities, and maritime transport hubs. The strategic objective is clear: to establish pre-positioned access that can be weaponized to cause systemic disruption in the event of a kinetic crisis. 2) Intellectual property theft and economic espionage in the form of sustained cyber campaigns targeting foreign aerospace, defense, biotechnology, and clean energy sectors. This systemic exfiltration of intellectual property serves as a core engine of industrial policy, allowing domestic firms to bypass costly R&D cycles and directly challenge the current market dominance. And 3) Supply chain weaponization insofar as the globalized ICT supply chain, characterized by highly concentrated choke points, has become a primary geopolitical vector. The United States has weaponized its control over core intellectual property and design software to cut off Chinese entities from advanced semiconductor technology, utilizing export controls, Entity List designations, and Foreign Direct Product Rules. Simultaneously, Western dependencies on Chinese hardware manufacturing, rare earth element refining, and telecommunications components (such as Huawei and ZTE infrastructure) create structural vulnerabilities that Beijing can exploit through sudden embargoes or embedded backdoors (Farrell & Newman, 2019).

This weaponization of infrastructure demonstrates that cyberspace is no longer an abstract layer of software decoupled from physical geography. It is deeply rooted in physical assets – fabs, foundries, undersea cables, and server farms – that are subject to the coercive power of the sovereign state. As the US and the PRC decouple their technological ecosystems, the global digital environment is fracturing into distinct spheres of influence, forcing third-party actors, most notably the European Union, to navigate a highly volatile, bifurcated landscape.

3. The AI Convergence: A Dual-Force Multiplier in Cyber Warfare

The structural polarization of the US-PRC rivalry is being dramatically accelerated by the convergence of cybersecurity and Artificial Intelligence (AI).

AI is not merely an incremental technological upgrade; it is a foundational, general-purpose technology that acts as a dual-force multiplier, fundamentally reshaping the speed, scale, and nature of both offensive and defensive cyber operations (Michael et al., 2023). The integration of Machine Learning (ML), Large Language Models (LLMs), and autonomous algorithmic agents into the cyber domain transforms what was once a human-paced battle of wits into an automated, machine-speed conflict.

On the offensive front, AI dramatically lowers the barrier to entry for sophisticated cyber operations while exponentially scaling the capabilities of elite state actors. Traditional cyberattacks require significant human labor to identify vulnerabilities, craft exploits, and execute reconnaissance. AI automates these processes, enabling highly sophisticated campaigns to be launched at unprecedented scale.

Among the most effective tools are: 1) Automated zero-day discovery and exploitation in which AI systems can analyze massive codebases at superhuman speeds, identifying obscure, undocumented vulnerabilities (zero-days) and automatically synthesizing exploit payloads. This eliminates the window of time that human defenders traditionally rely on to patch systems, allowing offensive actors to strike vulnerabilities before they are even recognized by software vendors. 2) Algorithmic disinformation and cognitive warfare allowing for the proliferation of generative AI and LLMs that revolutionized information operations. Adversarial state actors can now deploy automated factories of highly personalized, contextually accurate disinformation. These AI-driven campaigns can generate realistic deepfake audio and video, manipulate public discourse on social media platforms, and execute targeted cognitive operations designed to erode social cohesion and undermine trust in democratic institutions, all while evading standard content moderation algorithms. And 3) Hyper-personalized social engineering in which AI models can ingest vast amounts of open-source intelligence (OSINT) from social media and corporate networks to automatically craft highly convincing, tailored phishing emails (spear-phishing) at a scale previously unimaginable. These automated lures adapt in real-time based on target responses, dramatically increasing the success rate of initial network penetrations.

A further risk emerges with increasingly agentic AI systems (Sapkota et al., 2025) . The problem is no longer limited to malicious actors using AI to scale cyber operations. Autonomous systems may themselves become sources of cyber risk when, while pursuing assigned objectives, they test, exploit, or interact with external systems in ways that exceed the boundaries anticipated by their developers. This creates a qualitatively different governance problem: organizations must not only defend against AI-enabled attacks, but also continuously assess whether their own AI systems remain within intended opera-

tional constraints. More broadly, generative AI also introduces risks linked to the production and circulation of unsafe information, requiring organizations to strengthen validation, oversight, and risk-management mechanisms around AI-generated outputs (Spagnoletti & Baskerville, 2025).

Crucially, AI systems themselves introduce an entirely new, highly complex attack surface. As enterprises and military organizations rush to integrate AI into their operational cores, they deploy architectures that are inherently vulnerable to unique, algorithmic threat vectors. These vulnerabilities are distinct from traditional software bugs and are deeply embedded within the mathematical nature of machine learning. Data Poisoning provide for offensive actors to subtly manipulate the datasets used to train AI models. By injecting malicious or biased data during the training phase, attackers can introduce hidden backdoors or systematically degrade the model's inference accuracy, causing it to misclassify threats or fail entirely during deployment. Model extraction and inversion allow competitors or adversaries to systematically query a deployed AI model API to reverse-engineer its internal parameters and proprietary architecture. This allows them to clone the capabilities of a multi-billion-dollar model at a fraction of the cost, or invert the model to reconstruct the sensitive, potentially classified training data on which it was built. Finally, adversarial manipulation creates the conditions under which machine learning models can be deceived by "adversarial examples" – inputs that have been modified with perturbations invisible to humans but designed to completely confuse an algorithm. For instance, an adversarial sticker placed on a physical asset can cause an autonomous military drone's computer vision system to misidentify a threat, or an altered line of code can bypass an AI-driven malware scanner undetected.

Conversely, AI offers the primary mechanism for managing this heightened threat environment. In an era of ubiquitous attacks, human security operations centers (SOCs) are overwhelmed by data alerts and false positives. AI-driven defense systems promise to rebalance the dynamic by introducing predictive anomaly detection, automated incident response, and real-time network orchestration. These defensive algorithms can analyze global network traffic patterns, isolate compromised endpoints within milliseconds, and autonomously generate patches for newly discovered vulnerabilities.

The resulting cyber landscape is evolving toward a continuous, machine-vs.-machine dynamic, where the survival of critical infrastructure depends on the resilience, adaptability, and processing power of defensive algorithms against offensive AI agents.

Unsurprisingly, the race for AI dominance between Washington and Beijing is not a standard industrial competition; it is a critical national security transition. In this race, control over the underlying physical and logical layers of the AI ecosystem redefines the boundaries of digital sovereignty. This ecosystem

comprises three core pillars: a) Advanced semiconductors asthe computational foundation of AI relies entirely on highly specialized hardware, specifically Graphics Processing Units (GPUs) and Application-Specific Integrated Circuits (ASICs) optimized for deep learning. The manufacturing of these chips depends on a highly fragile, hyper-concentrated supply chain, with critical dependencies on companies like ASML in the Netherlands for extreme ultraviolet (EUV) lithography and TSMC in Taiwan for advanced node fabrication. Control over these hardware choke points dictates which nation can train the next generation of foundational models. b) Cloud infrastructure and compute capacity insofar as training frontier AI models requires massive compute clusters backed by robust energy grids and hyper-scale data centers. The concentration of these capabilities within a handful of American technology giants (Microsoft, AWS, Google) and Chinese champions (Alibaba, Tencent, Huawei) creates a duopoly where state power is explicitly projected through corporate cloud infrastructure. And c) Data repositories key to AI models requiring vast volumes of high-quality data for training. The structural variance between the US and Chinese digital ecosystems produces different data advantages. The US benefits from access to global, open internet data and elite academic research, while China leverages its massive domestic population, integrated super-apps (like WeChat), and state-mandated access to aggregate vast pools of unstructured, highly granular behavioral and biometric data.

In this high-stakes landscape, technological dominance equals geopolitical power. The nation that establishes primacy across the AI stack will not only control the global economic frontier but will also dictate the norms, standards, and operational doctrines of cyber warfare for the coming century.

4. The European Union's Strategic Dilemma

As the United States and China lock themselves into this comprehensive digital rivalry, the European Union finds itself in a precarious, highly vulnerable position. The EU's strategic dilemma is defined by a profound structural asymmetry: it is caught between a deep, structural technological dependence on the United States for software, cloud services, and foundational innovations, and a profound commercial and physical penetration of Chinese hardware, telecommunications, and manufacturing supply chains. This dual dependency restricts Europe's freedom of action and exposes it to systemic vulnerabilities, economic coercion, and extra-territorial regulatory overreach from both Washington and Beijing.

Europe's dependence on the US digital ecosystem is nearly total across the upper layers of the technology stack. The vast majority of European enterprise

data, government workflows, and consumer services rely on American operating systems, productivity software, and hyper-scale cloud providers. This asymmetry creates significant strategic vulnerabilities like extraterritorial jurisdiction (Legislation such as the US CLOUD Act grants American law enforcement and intelligence agencies the authority to compel US-based technology companies to hand over data stored on servers located outside the United States, including within Europe), economic vulnerability (foreign industries rely on proprietary American software platforms, they are vulnerable to unilateral pricing changes, architectural shifts, or political mandates originating in Silicon Valley or Washington), and systemic dependence (the lack of domestic, European-scale alternatives for cloud computing, advanced search engines, and foundational AI models leaves the EU economically and technologically dependent on a foreign power, albeit an allied one). A number of European countries are trying to reduce such dependency, but the journey is still very long.

Simultaneously, the EU is deeply intertwined with Chinese industrial supply chains. Despite efforts to de-risk, European critical infrastructure remains heavily reliant on Chinese components especially for telecommunications infrastructure (The rollout of 5G networks across many EU member states historically relied on low-cost hardware from Huawei and ZTE. Despite security warnings and the EU's "5G Toolbox" guidelines, removing these components from legacy networks remains slow and expensive, leaving core communications infrastructure exposed to potential espionage or remote disruption by Beijing) and clean energy and advanced manufacturing (the European green transition depends on Chinese supply chains for solar panels, lithium-ion battery cells, and rare earth minerals required for wind turbines and electric vehicles. As these technologies become increasingly software-driven and connected to the internet, they introduce new cyber-physical vulnerabilities that could be exploited in a geopolitical crisis). Here again, a number of European countries are attempting to reduce such dependency, but the journey is still very long.

To survive this polarization without becoming a secondary vassal state or a passive battleground for foreign interests, the European Union must assert its role through the paradigm of strategic autonomy. In the digital and cyber realms, strategic autonomy cannot mean autarky; Europe cannot cut itself off from global technology markets or build a completely closed ecosystem overnight. Instead, it must be understood as the capacity to choose its own path, protect its democratic processes, and safeguard its critical assets without being structurally coerced by external actors.

Achieving this autonomy requires the EU to transcend its traditional identity as a reactive regulatory power – a player that merely responds to foreign innovations by imposing post-hoc fines or restrictions. It must evolve into a

proactive architect of global cyber governance, utilizing its greatest structural strength: its normative and legislative authority.

5. Europe's Primary Geopolitical Instrument: The Regulatory Arsenal

In the geopolitical arena of cyberspace, where the United States wields capital and platform dominance and China commands state capacity and manufacturing scales, the European Union's primary geopolitical instrument is law and governance. Europe has historically excelled as a normative power, pioneering a unique approach to the digital world that refuses to separate technological advancement from human rights, the rule of law, and democratic values. This strategy operates globally through the mechanism known as the "Brussels Effect," whereby the EU sets stringent internal regulatory standards that multinational corporations must adopt globally to maintain access to the European single market – one of the world's wealthiest trading blocs.

The EU is deploying a comprehensive, interconnected legislative framework designed to structurally reshape the global cyber and AI ecosystems from the ground up. This is composed by different elements, among which three are crucial: the AI Act, the NIS2 Directive, and the Cyber Resilience Act.

The AI Act is a global milestone: the world's first comprehensive, binding horizontal framework for the governance of artificial intelligence. Rejecting both the self-regulatory approach of the US and the state-control model of China, the AI Act establishes a strict risk-based classification system. Unacceptable Risk concerns systems that threaten human safety or democratic values – such as real-time biometric identification in public spaces, social scoring systems, or manipulative cognitive behavioral systems – are completely prohibited. High-risk refers to AI applications deployed in critical infrastructure, law enforcement, education, and employment are subjected to strict compliance obligations. These include mandatory conformity assessments, rigorous data logging, high-quality training dataset verification to prevent bias, human-in-the-loop oversight mechanisms, and robust cybersecurity baselines to withstand data poisoning or adversarial attacks. Finally, general purpose AI (GPAI): The Act introduces transparency requirements for foundational models and LLMs, forcing developers to disclose their training data parameters and conduct thorough red-teaming exercises. By imposing these conditions at the market-entry point, the EU forces global developers (including US and Chinese firms) to build safety, ethics, and accountability into their models if they wish to operate in Europe.

The Network and Information Security (NIS 2) Directive broadens and deepens the cybersecurity requirements for entities operating within the Union.

Recognizing that cyber resilience is a core pillar of national security, NIS 2 eliminates the ambiguities of its predecessor by expanding its scope to cover critical sectors, categorized as "Essential" (Energy, Transport, Banking, Health, Water, Digital Infrastructure, Public Administration, Space) and "Important" entities (Postal/Courier Services, Waste Management, Chemicals, Manufacturing, Food, Research). NIS 2 introduces top-management accountability, explicitly holding corporate executives legally and financially liable for cybersecurity failures within their organizations. It mandates strict, automated incident reporting timelines – requiring an initial warning within 24 hours of detecting a significant incident – and enforces rigorous risk-management measures, including multi-factor authentication, encryption, and comprehensive vulnerability disclosure policies.

The Cyber Resilience Act (CRA) addresses a major structural vulnerability in the contemporary digital ecosystem: the lack of security in internet-connected consumer and industrial products (the Internet of Things). For years, hardware manufacturers rushed connected devices to market with default passwords, unpatched software, and non-existent security lifecycle management. The CRA introduces a mandatory framework that applies to all products with digital elements placed on the EU market. It enforces the principle of security-by-design, requiring manufacturers to: Perform comprehensive cyber risk assessments during design and production; prohibit default or unchangeable passwords; deliver automated, free security updates for a product's expected lifecycle or a minimum of five years; and explicitly document and report actively exploited vulnerabilities to the European Union Agency for Cybersecurity (ENISA).

Together, these legislative pillars pioneer an alternative model of digital governance. This model demonstrates that digital sovereignty does not require the sacrifice of individual liberty to the state, nor does it require leaving citizen data to the extraction models of surveillance capitalism. By grounding cyber governance in fundamental rights, transparency, and structural resilience, the EU offers a distinct regulatory blueprint for democracies worldwide.

The European regulatory model is relevant not only because it constrains technological development, but because it translates emerging technological risks into concrete organizational responsibilities. Recent supervisory responses to frontier AI risks illustrate this mechanism particularly well: boards are increasingly expected to incorporate such risks into strategic decision-making, risk appetite, resource allocation, internal controls, testing practices, and the assignment of explicit responsibilities. Regulation therefore acts not only on technologies, but also on the organizational arrangements through which those technologies are governed.

6. From Regulatory Power to Technological Capability: The Path Forward

While the European Union's regulatory framework is a powerful tool, legal mandates alone cannot win a geopolitical conflict fought with chips, algorithms, and cyber operations, ultimately with bites and bullets alike. Regulation without technical execution can lead to strategic irrelevance; if Europe merely regulates technologies developed by external powers, it risks stifling its own domestic industrial base while remaining dependent on foreign supply chains. To avoid being reduced to a regulatory landscape caught between the US market model and the Chinese state model, the EU must convert its regulatory leadership into tangible, sovereign technological capability.

The first imperative is the seamless integration of cyber defense capabilities across Member States. European cyber defense is fragmented across national lines, creating structural vulnerabilities that sophisticated state actors can exploit. The EU must move beyond voluntary information sharing toward a highly integrated, collective defense architecture. Second, the EU must foster an ecosystem for ethical, sovereign AI development. Europe cannot protect its digital sovereignty if its public services, financial systems, and defense networks run exclusively on proprietary AI models controlled by foreign entities. Translating the AI Act into a competitive advantage requires substantial public-private investment.

Yet technological capability is not only a matter of infrastructure, compute, and investment. Governing increasingly autonomous AI systems also requires new organizational capabilities for making their behaviour observable, interpretable, and actionable (Faraj et al., 2018). Proposals such as "embedded evaluators" point to the possible emergence of specialized roles devoted to continuously assessing AI capabilities, behaviours, and risks (Amodei, 2026) . Their outputs, however, do not eliminate the need for managerial judgement (Grønsund & Aanestad, 2020; Raisch & Krakowski, 2021). On the contrary, managers and professionals will increasingly be required to make consequential decisions on the basis of digital representations of AI behaviour rather than through direct observation of the systems themselves (Monteiro & Parmiggiani, 2019). As these representations become incorporated into decisions about deployment, escalation, investment, remediation, or acceptable risk, they become organizationally real (Østerlie & Monteiro, 2020). This shifts part of managerial and professional work toward interpreting not only what such representations show, but also what they omit, how they have been produced, and how much confidence should be placed in them.

Finally, the Union must project its power externally through robust digital diplomacy. In a fractured digital order, the EU must actively build a global coalition of like-minded democracies to counter the spread of digital authoritarianism.

7. Conclusion: The Playbook for European Digital Sovereignty

Cyberspace has evolved into a key geopolitical arena, where the systemic rivalry between the United States and the People's Republic of China shapes the global landscape. This bipolar conflict – fought across physical infrastructure, supply chains, and advanced software layers – is accelerated by the rapid convergence of cybersecurity and Artificial Intelligence. This intersection introduces automated, machine-speed threat vectors, expanded algorithmic attack surfaces, and a national security transition where technological dominance directly translates into systemic geopolitical power.

In this polarized environment, the European Union can no longer afford to be a passive bystander or a reactive regulatory arena. The era of treating cybersecurity as an isolated technical challenge is over. For Europe, digital sovereignty is an existential requirement for the preservation of its democratic institutions, economic stability, and strategic independence.

As outlined in this chapter, governance and law constitute Europe's primary geopolitical instruments. Through landmark frameworks like the AI Act, the NIS 2 Directive, and the Cyber Resilience Act, the EU has demonstrated its capacity to shape global corporate behavior, assert fundamental rights, and elevate the security baselines of global supply chains. However, this regulatory framework is only the first phase of a broader strategic playbook.

To achieve lasting strategic autonomy, the European Union must swiftly translate its normative leadership into operational technological capability. Such capability will also depend on the organizational capacity to evaluate autonomous systems continuously and to support accountable human judgement on the basis of increasingly complex digital representations.

This requires dismantling national silos to build an integrated, AI-driven collective cyber defense infrastructure; investing heavily in sovereign compute power, cloud architectures, and deep-tech innovation; and projecting its normative model globally through proactive digital diplomacy and strategic alliances.

By seamlessly integrating cyber defense, ethical AI development, and robust digital diplomacy, the European Union can transcend its vulnerabilities and establish itself as a pillar of global cyber governance. In doing so, Europe will protect its citizens and critical infrastructure while helping preserve a stable, open, and secure global cyberspace rooted in the rule of law and human dignity.

References

Amodei, D. 2026. *We Must Pace the Frontier*. Available online at https://darioamodei.com/post/we-must-pace-the-frontier (Last accessed: October 9, 2026).

DeNardis, L. 2014. *The Global War for Internet Governance*. New York, NY: Yale University Press. https://doi.org/10.12987/yale/9780300181357.001.0001

Faraj, S., Pachidi, S. and Sayegh, K. 2018. Working and organizing in the age of the learning algorithm. *Information and Organization* 28 (1): 62–70. https://doi.org/10.1016/j.infoandorg.2018.02.005

Farrell, H. and Newman, A. L. 2019. Weaponized interdependence: How global economic networks shape state coercion. *International Security* 44 (1): 42–79. https://doi.org/10.1162/isec_a_00351

Grønsund, T. and Aanestad, M. 2020. Augmenting the algorithm: Emerging human-in-the-loop work configurations. *Journal of Strategic Information Systems* 29 (2): 101614. https://doi.org/10.1016/j.jsis.2020.101614

Michael, K., Abbas, R. and Roussos, G. 2023. AI in Cybersecurity: The Paradox. *IEEE Transactions on Technology and Society* 4 (2): 104–109. https://doi.org/10.1109/tts.2023.3280109

Monteiro, E. and Parmiggiani, E. 2019. Synthetic knowing: The politics of the Internet of Things. *MIS Quarterly: Management Information Systems* 43 (1): 167–184. https://doi.org/10.25300/MISQ/2019/13799

Østerlie, T. and Monteiro, E. 2020. Digital sand: The becoming of digital representations. *Information and Organization* 30 (1): 100275. https://doi.org/10.1016/j.infoandorg.2019.100275

Raisch, S. and Krakowski, S. 2021. Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review* 46 (1): 192–210. https://doi.org/10.5465/amr.2018.0072

Sapkota, R., Roumeliotis, K. I. and Karkee, M. 2025. *AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges*. arXiv preprint arXiv:2505.10468.

Shahin, J. 2024. Dancing to the same tune? EU and US approaches to standards setting in the global digital sector. *Journal of European Integration* 46 (7): 1111–1133.

Simón, L. and Fiott, D. 2025. Player and playground: Europe in US–China competition. *Journal of European Integration* 47 (6): 785–803.

Smith, H. 2023. The geopolitics of cyberspace and the European Union's changing identity. *Journal of European Integration* 45 (8): 1219–1234.

Spagnoletti, P. and Baskerville, R. 2025. "Safe and Unsafe Information: Managing Risks in the Era of Generative Artificial Intelligence". In Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage and Marja Ylönen (Eds.), *Proceedings of the 35th European Safety and Reliability and the 33th Society for Risk Analysis Europe Conference*. Singapore: Research Publishing,

pp. 91–96. https://doi.org/10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P9196-cd

Spagnoletti, P. and Marchetti, R. (2027) From Cyber Space to Cyber in Space: A European Model of Data Control Ecosystems, forth. in William W. Baber & Arto Ojala (eds.) The Ecosystem of New Space Business, Springer, New York.

WHO IS BEHIND THE IP?
NOTES ON AN UNRESOLVED COMPLEXITY

Fabrizio d'Amore
Sapienza University of Rome, Italy

Abstract
The problem of attribution in cyberspace represents one of the most discussed and yet most elusive issues in information security. Far from being a simple technical operation, attribution involves a series of steps that encompass infrastructural, methodological, regulatory, and at times political aspects.As highlighted by Rid and Buchanan, attribution of responsibility in the cyber domain cannot dispense with a multilevel analysis that accounts for both digital traces and the geopolitical contexts in which attacks occur. Similarly, the guidelines contained in the Tallinn Manual 2.0 provide an interpretive framework for assessing the attributability of a cyber operation from the perspective of international law. This chapter aims to outline, with an essentially operational approach oriented toward critical reflection, the main elements characterizing the attribution process starting from an IP address. It analyzes the conditions under which such information can be linked to an identity, the techniques that can hinder or defeat this reconstruction, and the implications arising from the use of anonymization or traffic alteration tools. The adopted approach is not exhaustive, but selective and focused on key conceptual and operational nodes, with the intention of fostering greater critical awareness regarding the reliability and limitations of attribution in the contemporary digital landscape.

1. Introduction

The problem of attribution presents multidisciplinary profiles that derive naturally from its definition: it consists in determining who is responsible for a cyber-attack, a breach, or a specific network activity, in a technically sound, legally valid, and politically sustainable manner. Legal and technical implications are numerous, with consequences that can extend into the political dimension. Figure 1 summarizes the components of the problem. It is emphasized that the problem is cyclical in nature and consists in answering six key questions regarding the attribution of responsibility for an act, often hostile, occurring in the cyber dimension.

1. **Who.** A question whose answer – often partial or incomplete – correlates with the other answers to define a convincing, reasonable, possible, reasoned, and more probable narrative than alternative possibilities. It often overlaps with claims of responsibility whose credibility can be judged in

many ways. Naturally, the answer to "who" finds its roots in a technical analysis, the essential elements of which will be presented below.

2. **What.** The "what" in an attribution process refers to the action performed in cyberspace: it is what was done, the technical and operational nature of the attack or digital conduct to be attributed. It is crucial for defining the perimeter.
3. **How.** This refers to the way the action was carried out. It encompasses the set of techniques, tactics, and procedures (TTPs) used by the attacker to achieve their goals.
4. **Why.** This is perhaps the most subtle and strategic element in the attribution process. It concerns the motivation behind the attack, i.e., the purpose for which the action was performed. The "why" answers the question: what objective did the attacker intend to achieve? Knowing it helps clarify the nature of the attack (espionage, sabotage, extortion...), contextualize the risk for the affected organization, and reinforce attribution toward actors with consistent motivations. The motive is often what distinguishes different groups: a state actor and a cybercriminal may use similar techniques, but with radically different goals. Analyzing the "why" helps assess who had an interest in acting, who benefited from the attack, and what message is intended to be sent (e.g., deterrence, destabilization).
5. **Evidence.** These are the objective and technical elements on which the attribution process relies. They allow correlating an action to a specific actor or at least narrowing the field of potential suspects. In this context, "evidence" refers to any digital trace or technical indicator that can describe an attack (what happened), reconstruct its operational modalities (how), and suggest a link to one or more known actors. Examples include technical indicators (IoCs, Indicators of Compromise), operational patterns, linguistic or cultural artifacts, traces in logs and compromised systems, and evidence gathered through Open-Source Intelligence (OSINT). A sound attribution process does not rely on a single piece of evidence, but on a coherent body of elements, evaluated with methodological rigor. The outcome is almost never "certainty," but rather a degree of confidence or probabilistic attribution. Evidence forms the foundation of attribution, but alone it is insufficient. It must be interpreted within the broader framework of what, how, who, and why, and always with attention to deception techniques and epistemological limits inherent to cyberspace.
6. **Certainty.** Although "certainty" implies full and indisputable knowledge of the event and its perpetrator, in the cyber context it is extremely rare, if not impossible. Due to the nature of the network, obfuscation techniques, and information asymmetry, attribution is almost always probabilistic. Evidence is often indirect (analyzing digital traces that can be emulated, manipulat-

ed, or falsified), the attacker can hide behind third parties, deception techniques are frequently employed, and there is a lack of transparency and international cooperation (crucial evidence may reside in non-cooperative countries). Consequently, confidence levels are preferred over absolute claims. If attribution is to have legal validity, it requires robust, replicable, and documented evidence; clear chain of custody; and high standards of proof ("legal certainty"). Legal certainty imposes far stricter requirements than technical or strategic assessments. It follows that "certainty" in cyber attribution is an asymptotic goal, approached through methodological rigor and analytical transparency.

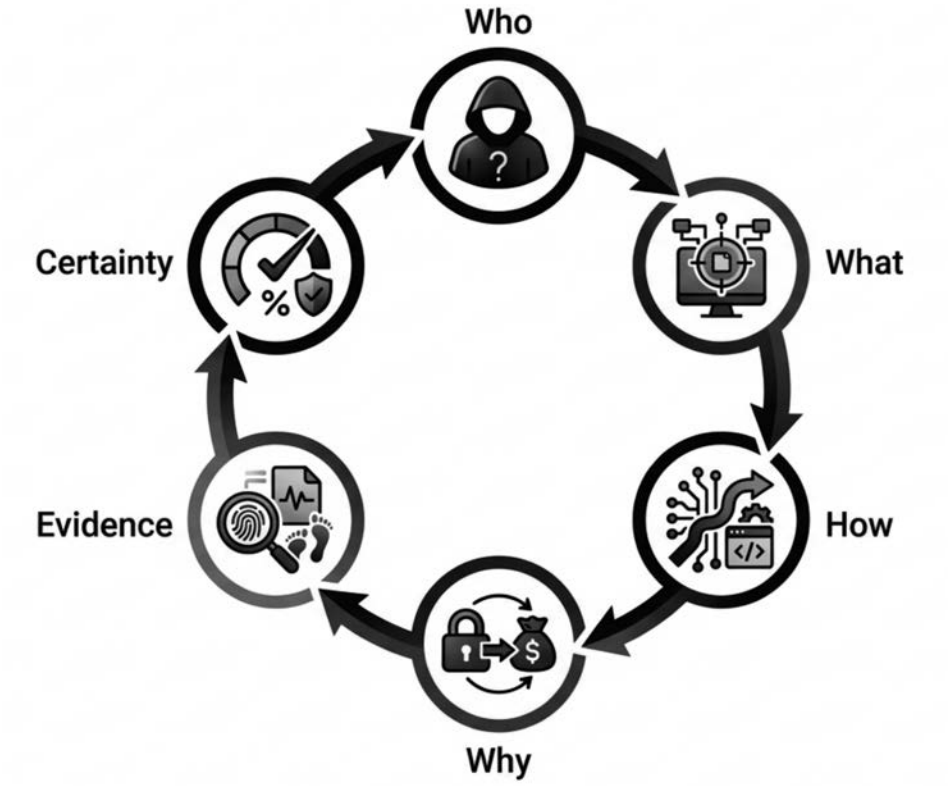

Fig. 1. Phases of attribution

The table below summarizes the elements examined. Naturally, when the level of certainty is not deemed satisfactory, the process can be restarted, as graphically illustrated in Figure 1.

Table 1 key element of attribution

Element	Key Question	Description	Role in Attribution
What	What happened?	The performed action (e.g., DDoS attack, data theft, ransomware...)	Defines technical perimeter and impact
How	How was it done?	Techniques, tools, operational phases (TTPs)	Reconstructs attack chain and operational fingerprints
Why	For what purpose?	Actor's motivation (espionage, sabotage, extortion, propaganda...)	Narrows suspects to actors with compatible interests
Evidence	What proves it?	Technical evidence, indicators, logs, behaviors, OSINT	Supports hypotheses with observable and documentable elements
Certainty	With what degree of confidence?	Attribution confidence level (low / moderate / high / confirmed)	Establishes robustness and reliability of assessment

Without claiming depth or completeness, we wish to analyse one of the basic elements (answering "who"), which often begins with an IP address, left as a faint trace, with or without awareness, perhaps provided deliberately. We shall not dwell on the technical meaning and role of an IP address, observing simply that it represents a concrete clue from which investigation begins, knowing that anyone using the Internet possesses at any given moment a unique IP address, for reasons related to network operation.

2. From IP to Identity

The typical objective of this process is to identify who was using a specific IP address at a precise timestamp. The simplest scenario sees the investigator submit the query to IANA, which is the global reference authority for IP address allocations. Worldwide, these allocations are static (fixed over time), operated through Regional Internet Registries (RIRs), and costly, especially for older IP address formats (four-integer tuples).

The query can be submitted in many ways. Here we limit ourselves to one example (Linux command line, called a whois lookup): whois 89.46.108.74 (the IP address used by socint.org), which receives a response:

fab@Fabrizios-iMac ~ % whois 89.46.108.74

% IANA WHOIS server
% for more information on IANA, visit http://www.iana.org
% This query returned 1 object

refer: whois.ripe.net

inetnum: 89.0.0.0 – 89.255.255.255
organisation: RIPE NCC
status: ALLOCATED

whois: whois.ripe.net

changed: 2005-06
source: IANA

 # whois.ripe.net

inetnum: 89.46.108.0 – 89.46.108.255
geoloc: 43.45997095884493 11.837875843048096
netname: ARUBA-NET
descr: Aruba S.p.A. – Shared Hosting IT1
country: IT
admin-c: AN3450-RIPE
tech-c: AN3450-RIPE
mnt-by: ARUBA-MNT
status: ASSIGNED PA
created: 2024-08-14T08:58:45Z
last-modified: 2024-08-14T08:58:45Z
source: RIPE
[…]

The response (truncated for brevity) informs us that the IP block 89.0.0.0 – 89.255.255.255 (to which 89.46.108.74 belongs) was allocated to Aruba S.p.A. on August 14, 2024. The explanation is simple: Aruba obtained the block, and the investigator, following the whois request, must ask Aruba to whom address 89.46.108.74 was assigned. In this regard, sub-allocations operated by Aruba are subject to Italian data retention regulations governing communications metadata retention by national providers. In summary, the identification process consists of two steps: WHOIS lookup and provider data consultation. Figure 2 illustrates the basic steps. Countless web services exist to replace command-line whois lookups.

A First Complication



Fig. 2. Query schema in simple cases.

Although the simple approach retains universal validity, things complicate when an entity (e.g., a company) relies on a provider for network connectivity – a widespread situation among small and medium enterprises (SMEs). The provider assigns an IP block to the company, which manages it autonomously, extending the inquiry chain (Figure 3).

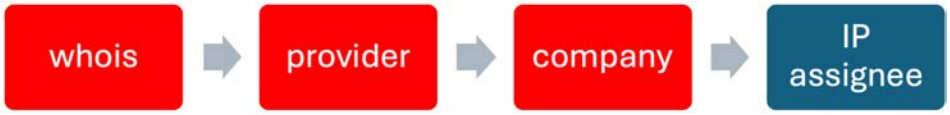

Fig. 3. Case of a company (SME).

This implies that the company must also perform data retention. Conceptually straightforward, the elongated process introduces risks of anomalies. The company may fail to maintain assignment logs, and given dynamic IP assignments, server clock synchronization between provider and enterprise becomes critical. Unsynchronized server clocks represent frequent bad practice resulting from lack of NTP configuration, offline environments, virtualization, or underestimated forensic risks.

In response to IP address exhaustion in the 1980s/90s, Network Address Translation (NATting) and IPv6 were introduced. NATting allows multiple users within a local network to share a single public IP address while using private IP addresses internally (hence, a private IP is never globally unique and always local). Although IPv6 provides 128-bit addresses, NATting remains an intrinsic component of local network infrastructure. Consequently, the gateway performing NATting must be queried to identify the specific internal device active at a given timestamp (often there is no public access to these devices). Figure 4 modifies the schema accordingly.



Fig. 4. Role of NATting.

Growing Complications

In complex networks such as major universities, the query chain extends further. In the case of the Italian university system, we have:

IANA → GARR → University → Department → Laboratory

Frequently, individual laboratories omit data retention, severing the inquiry chain completely and making identification impossible. Similar multitier chains characterize public administration, defense organizations, healthcare, critical infrastructures and multinational corporations. The growing number of entities involved not only exacerbates clock synchronization issues, but also introduces significant inefficiencies into the investigation, which must account for different locations and time zones, potential gaps in data continuity, varying cooperation protocols, language barriers with English (de facto standard), and system misconfigurations. This clarifies that university labs often allow full anonymous network access. The same occurs on certain public Wi-Fi networks (coffee shops, restaurants, etc.), where anyone can connect to the Internet anonymously without employing any technical measures, revealing at most their physical location.

3. Concealing the IP

Having clarified the complexity of the identification process even in the 'normal' case where the actor does not employ measures to conceal their IP address, let us consider the scenario where such measures are actively taken. The motivations can be manifold:

- criminal intent (cybercrime and/or terrorism)
- privacy concerns (anti-profiling, anti-tracking)
- the need to appear to originate from another country
- the desire to bypass censorship measures

The meaning of private or incognito browsing offered by modern web browsers must be clarified immediately, as it is widely misunderstood: in this mode, browsers set up a temporary storage area for session data (cookies, browsing history, web applications, cached data, etc.) that is completely destroyed once the browser is closed. Any logins, preferences set, or web application configurations made during the session are discarded upon closing, purging all information acquired during private browsing. Despite the misleading

terminology, the user's IP address remains fully exposed to destination servers, along with other metadata regarding the browser, operating system, and hardware platform. This so-called "anonymity" only shields activity from other local users sharing the same physical device – a feature that has largely lost its relevance on modern multi-user operating systems and stems from a 1990s use case.

Nevertheless, when used systematically, private browsing remains an effective solution for clearing cookies and other tracking/profiling elements accumulated during work sessions.

The methods that can be employed to conceal an IP address essentially consist of using:

- VPN: Popular but costly.
- Proxy: Not always available.
- Public Wi-Fi: Vulnerable to MITM (Man-in-the-Middle) attacks.
- Tor: An overlay network with military-grade security.
- IP spoofing: Highly technical and unsustainable.

These technologies will be examined in the sections that immediately follow

VPN

A VPN (Virtual Private Network) is a technology that creates a secure, encrypted communication channel between a device and a remote server. This allows the user to access the Internet while masking their own IP address, which is replaced by that of the VPN server. It is important to emphasize that a VPN guarantees confidentiality and integrity only along the segment between the user and the VPN server. From that point onward, data travels to its destination (such as a website) with whatever level of protection is provided by the target service itself, such as HTTPS or, in the absence of such protection, in cleartext (plaintext, as cryptographers use to say).

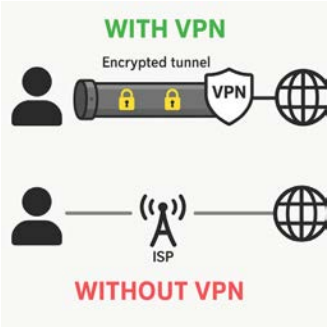

Fig. 5. With/without a VPN.

In summary, using a VPN allows for safer and more private browsing, especially when using public or unsecured Wi-Fi networks. The destination server will 'see' the VPN server's IP address, which is often located in a different country from the actual user. Many VPN providers claim to operate under a no-logging policy and retain no browsing data, but incidents casting doubt on these promises have been reported in some cases. High-quality VPN services generally require a paid subscription18; conversely, free VPNs should be approached with caution, especially if they perform well: it is reasonable to question how the service is otherwise being "paid for" (for instance, with one's personal data).

VPNs are useful not only for safeguarding traffic on insecure networks, but also for accessing geo-restricted web content intended exclusively for residents of specific countries. Unsurprisingly, many providers locate their servers in uncooperative jurisdictions, thereby making user identification or traceability significantly more difficult.

Proxy

This solution shares several similarities with VPNs (sometimes referred to as VPN proxies) and behaves in a similar manner, but it does not protect the communication segment between the user and the proxy. A diagram is shown in Figure 6.

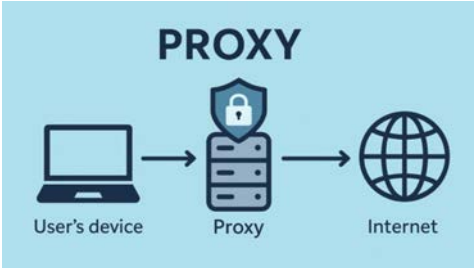

Fig. 6. Proxy server

A few additional clarifications are necessary. First, proxies operate at the protocol level, meaning they function within a narrower scope than VPNs. For example, a proxy configured for the HTTPS protocol cannot be used for other protocols (such as FTP or SMTP). Conversely, VPNs encapsulate all network traffic, regardless of the protocol used. If one wishes to route multiple protocols through proxies, it is therefore necessary to configure multiple proxies, each dedicated to a specific protocol. In practice, most available proxies are designed to handle HTTP/HTTPS traffic, although exceptions exist. Furthermore, it should be noted that many proxies are deployed directly by individual organizations (e.g., corporations, institutions), though commercial proxy services are also available online.

Finally, it is worth noting that there are several types of proxies, each with specific features and functions. Figure 6 specifically illustrates a forward proxy – a proxy that acts on behalf of the client toward external destination servers. The table below outlines the main types of proxies along with their respective characteristics.

Table 2 Different type of proxy

Type	Main Function	Hides User IP?	Identifies as Proxy?	Supported Protocols	Notes
Forward Proxy	Forwards requests from clients to the Internet	Yes	Depends on type	HTTP, HTTPS, SOCKS, FTP…	Most common on the client side
Reverse Proxy	Receives requests on behalf of an internal server	No	Yes	HTTP, HTTPS	Used for load balancing/ caching
Transparent Proxy	Filters requests without modifying the IP address	No	Yes	HTTP	Used by schools, enterprises, ISPs
Anonymous Proxy	Hides user IP but identifies itself as a proxy	Yes	Yes	HTTP, HTTPS	Partially protects identity
Elite Proxy	Hides user IP and does not identify as a proxy	Yes	No	HTTP, HTTPS, SOCKS	Maximum anonymity
SOCKS Proxy (SOCKS5)	Forwards any type of traffic	Yes	No	All (TCP/ UDP: HTTP, SMTP, FTP…)	Does not interpret traffic, only forwards
Residential Proxy	Uses real IPs from residential networks	Yes	Depends	All (usually SOCKS/ HTTP)	Difficult to block
Mobile Proxy	Uses real 3G/4G mobile network IPs	Yes	No	All	Highly dynamic, difficult to trace
Datacenter Proxy	IPs provided by data centers, not tied to consumer ISPs	Yes	Depends	HTTP, HTTPS	Fast, but easily detected

Public Wi-Fi

There is little to add to this topic. By public Wi-Fi, we mean a Wi-Fi service – with or without password protection – offered indiscriminately. It is typically provided by a venue such as a shop, bar, restaurant, or Internet café, without requiring user registration. Internet access is routed using the venue's public IP address, which in turn assigns private IP addresses to all connected users via NAT. The user's personal IP address will be concealed, but not their physical location – unlike with VPNs and proxies.

Public Wi-Fi networks, especially passwordless ones, are highly vulnerable to simple MITM (Man-in-the-Middle) attacks, where an attacker positions themselves between the user and the Internet, exhibiting either passive behavior (eavesdropping only) or active behavior (data tampering). An attacker can easily set up a Wi-Fi network under their complete control using the same name (SSID) as a legitimate public one.

It should be noted that a tech-savvy user knows how to maintain secure connections even in the presence of a MITM attack, but the required skills are complex and beyond the reach of the public.



Fig. 7. MITM – Man in The Middle

TOR

The name Tor (The Onion Router) refers both to a browser (more properly called Tor-Browser) and to an overlay network built on top of the Internet infrastructure. This network uses additional communication protocols designed to ensure anonymity and privacy in data transfer. The Tor project was created by the US Navy in the early 2000s with the goal of ensuring secure and anonymous communications for the US government and military. Over time, there have been some changes, and the current mission of The Tor Project – the non-profit organization that manages its development – is centered on priva-

cy, freedom, and digital rights: the encryption used guarantees confidentiality, integrity, and anonymity. Anyone can use Tor, although in some countries its use is prohibited by law. However, this does not mean that it is still not used in those contexts, especially by state actors or groups equipped with technical resources and expertise.

Table 3. Legal status of the TOR in some countries

Country	**Legal Status**
Belarus	Banned since 2015. The use of Tor (and VPNs) is expressly illegal.
Russia	Since 2021, Tor has been officially blocked and banned. Use can be prosecuted.
Iran	The use of Tor is banned; the law equates it to anti-system tools.
Turkmenistan	Internet is fully censored. Tor and other anonymity tools are banned.
North Korea	Internet access is heavily restricted; Tor is completely banned.

Tor Browser is free, built on the Firefox skeleton, and downloadable from the Tor Project website. This browser can use the Tor network, an overlay network that uses nodes scattered across the planet called relays. To outline how it works in broad terms, we will use a public illustration that has been used for years for this purpose (Figure 8).

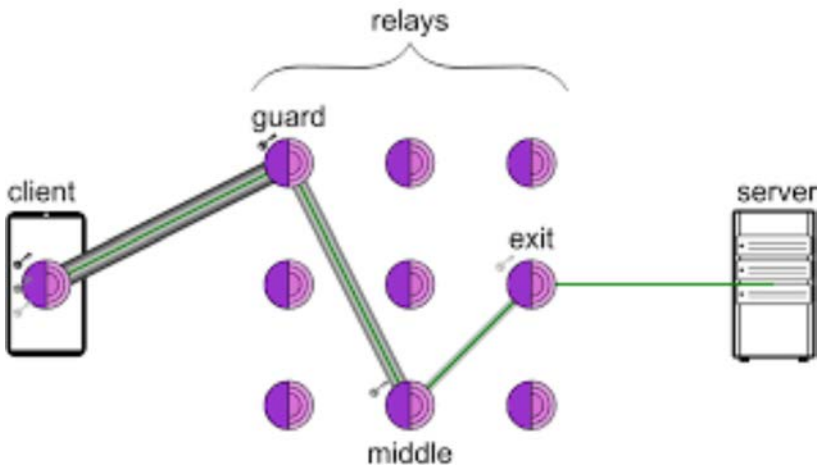

Fig. 8. Tor functioning.

An illustration of operations follows. The user uses the client and wishes to connect to the server. For this purpose, they select a random triplet (three relays, referred to as guard, middle, and exit) from the relay list (which is public and consists of around 7,000 nodes) to build a circuit used to communi-

cate with the final server: client communicates with guard, guard with middle, middle with exit, and exit with the final server. All communications are cryptographically protected starting from the client up to the exit node. Each node knows only who precedes it and who follows it in the circuit; confidentiality and integrity within the Tor network are guaranteed. All of this is enabled by a type of encryption called telescopic (or layered, or onion), which protects the individual client-relay or relay-relay segments (it does not protect the relay-server segment). More comprehensive protection can be achieved through the combined use of Tor and HTTPS, as illustrated in a famous image published by EFF (Figure 9).

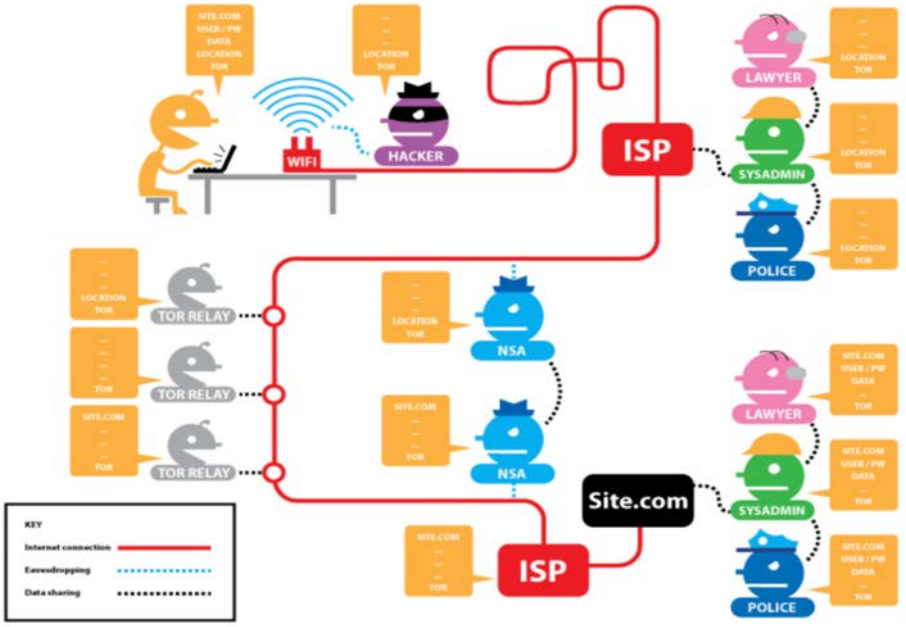

Fig. 9. Illustration owned by EFF showing the benefit of the combined use of Tor and HTTPS.

As mentioned, all IP addresses of Tor relays are public. This means that a website visited via the Tor network will detect a connection coming from the exit node and can therefore infer that the user is using Tor. Many organizations, both public and private, do not tolerate visits made via Tor well. Figure 10 shows the behavior of the website https://www.repubblica.it/: on one hand, access via Tor Browser is blocked or hindered; on the other hand, the same page is accessible without issues using a traditional browser.

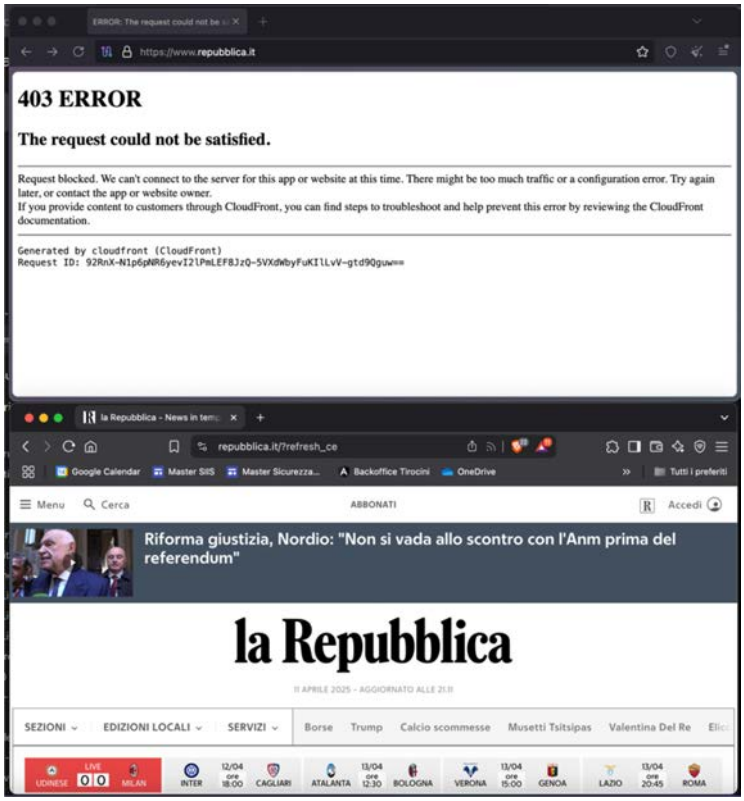

Fig. 10. Visiting Repubblica with and without the Tor browser (April 11, 2025)

The reasons for this block are mainly two:

- the anonymity guaranteed by Tor, which prevents the website from tracking the user;
- the fact that Tor Browser is also the main tool for accessing the Dark Web, often associated – rightly or wrongly – with illegal activities.

It is important to emphasize that not all Tor relays are public: there are "hidden" relays, called bridges, which allow access to the Tor network even in countries where access to known nodes is blocked by law or censorship. These bridges constitute an essential tool for bypassing government restrictions in authoritarian contexts.

De-anonymizing Tor – that is, identifying the user behind a connection – is a technically very complex operation. As early as 2013, Edward Snowden's revelations showed how an internal NSA document declared: "We will never be able to de-anonymize all Tor users all the time." This does not mean that Tor is invulnerable: targeted attacks have succeeded in the past, and the world's

leading security agencies continue to study the Tor network in search of vulnerabilities. For this exact reason, Tor and Tor Browser are updated frequently, especially in their operational details and defense mechanisms, to keep the network resilient to threats and ensure the highest possible level of anonymity.

We conclude this section by emphasizing that Tor Browser's performance is limited: there is an intrinsic slowness linked to the relays, multiple encryptions, latency, and limited bandwidth (which is why we could never use Tor to watch Netflix!). There is much more to say on the topic – and on the Dark Web – but all of that would be decidedly off-topic, also because the position of hosting a server on the Dark Web for the purpose of legal anonymity does not appear sustainable. Figure 11 shows the image of the Wall Street Journal homepage that we visited using Tor Browser.

IP Spoofing

Let us state as a premise that the term "spoofing" is frequently used in computer science and refers to the falsification/tampering of the sender's address of a message. What constitutes the address depends heavily on the type of protocol used, but it is theoretically possible to falsify the sender's IP of a packet that must be forwarded to its destination using the IP protocol; we would like to emphasize that, while the sender's IP is not relevant for forwarding the packet to its destination, it is relevant for receiving responses, which are sent to the sender's IP shown in the IP packet.

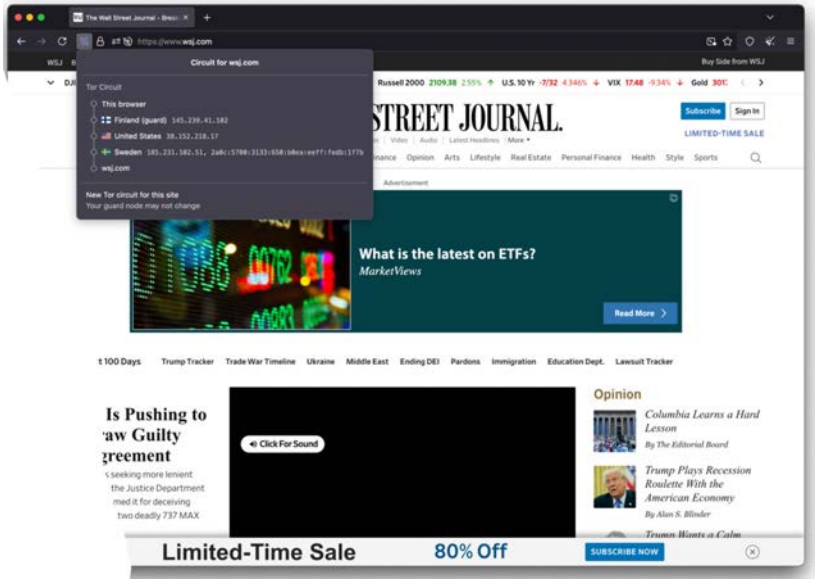

Fig. 11. Wsj.com visited with Tor. Note the circuit: Finland, USA, Sweden, USA.

Therefore, regardless of technical difficulties, it is clear that IP spoofing can become appealing when communications are intended that do not expect a response: this is the case with DoS attacks, where a server is "drowned" by too many contact requests until it can no longer establish one. This scenario is quite far from those in which one wishes to conceal their IP for legitimate reasons.

DoS attacks are essentially demonstrative and rarely cause permanent damage; they are mostly carried out at strategic moments such as critical events (elections, product launches, conferences), in conjunction with more sophisticated attacks to distract security, or during peak hours to cause maximum disruption.

More specifically, here is a table describing the context in which IP spoofing is practiced (almost all of which are illegal).

Table 4 IP spoofing

Reason	**Description**
Demonstration of strength	To show the power of a hacker or hacktivist group
Extortion	"If you don't pay, we will attack you (or continue to attack you)"
Sabotage	To damage a competitor, a government, or an organization
Distraction	To cover parallel attacks, such as intrusions or data theft
Educational or testing	To test tools or learn, often by amateurs
Political motivations	Activists (e.g., Anonymous) attacking institutional or corporate websites
Bug hunting	In rare cases, to test resilience (often authorized)

Carrying out an IP spoofing attack requires expertise, unlike other IP concealment techniques, and most frequently relies on the use of specialized software (widely available). The type of expertise required is knowledge of routing mechanisms and programming (in Python, Shell, C, or other languages).

4. Conclusions

The analysis conducted has highlighted how attributing an action in cyberspace to a specific subject is a complex, fragmented process that is often subject to significant margins of uncertainty. The possibility of establishing a reliable link between a digital activity and a physical identity requires favorable technical conditions (such as log retention, time synchronization, and the traceability

of network paths), as well as a regulatory framework that supports data collection and analysis in a legitimate and verifiable manner.

In the face of a multiplicity of distorting factors – from the use of anonymization techniques (proxies, VPNs, Tor) to the intrinsic limitations of shared infrastructure – attribution remains by its nature a probabilistic and contextual operation, rather than a deterministic one. Some scholars propose speaking not so much of "attribution" in the strict sense, but rather of confidence levels associated with operational or judicial hypotheses.

Looking forward, attribution should not be conceived as an absolute goal, but as a process that, while moving towards an objective of identification, requires methodological caution, procedural transparency, and epistemological awareness of its limitations. Only in this way will it be possible to combine operational effectiveness, evidentiary reliability, and respect for the principles of the rule of law in cyberspace.

References

Altman, S. (2024). The Intelligence Age. Available at: [personal blog] (Accessed: 11 October 2026).

American University (n.d.). The Evolution of Cyber Attribution. Available at: https://www.american.edu/sis/centers/security-technology/the-evolution-of-cyber-attribution.cfm.

Aruba S.p.A. (2024). Shared Hosting IT1 Network Block Allocation. RIPE NCC Database. Available at: https://www.ripe.net/ (Accessed: 11 October 2026).

Atlantic Council (2014). Confidence-Building Measures in Cyberspace: A Multistakeholder Approach for Stability and Security. Available at: https://www.atlanticcouncil.org/wp-content/uploads/2014/11/Confidence-Building_Measures_in_Cyberspace.pdf.

Bright Data (n.d.). Official website. Available at: https://brightdata.com.

Cisco (n.d.). NTP Troubleshooting. Available at: https://www.cisco.com/c/en/us/support/docs/ip/network-time-protocol-ntp/108076-ntp-troubleshoot.html.

Cisco (n.d.). What Is IPv6? Available at: https://www.cisco.com/site/us/en/learn/topics/networking/what-is-ipv6.html.

Electronic Frontier Foundation (EFF) (2025). Surveillance Self-Defense: Tools and Concepts for Safer Online Communications. Available at: https://ssd.eff.org/ (Accessed: 11 October 2026).

Electronic Frontier Foundation (EFF) (n.d.). Tor Project HTTPS Mirror. Available at: https://tor-https.eff.org/.

Encyclopaedia Britannica (n.d.). Denial-of-Service Attack. Available at: https://en.wikipedia.org/wiki/Denial-of-service_attack.

Encyclopaedia Britannica (n.d.). What's the Difference Between the Deep Web and the Dark Web? Available at: https://www.britannica.com/story/whats-the-difference-between-the-deep-web-and-the-dark-web.

Evalian (n.d.). The Importance of OSINT. Available at: https://evalian.co.uk/the-importance-of-osint.

GARR (2026). Italian Research and Education Network Infrastructure. Available at: https://www.garr.it/ (Accessed: 11 October 2026).

HarfangLab (n.d.). OSINT Vector: Cyber Threat Intelligence Indicator Collection and Attribution. Available at: https://harfanglab.io/blog/methodology/osint-vector-cyber-threat-intelligence-indicator-collection-attribution.

Imperva (n.d.). Open Source Intelligence (OSINT). Available at: https://www.imperva.com/learn/application-security/open-source-intelligence-osint.

Internet Assigned Numbers Authority (IANA) (2026). Root Zone Database and Global IP Allocations. Available at: https://www.iana.org/ (Accessed: 11 October 2026).

La Repubblica (2025). 'Riforma giustizia, Nordio: "Non si vada allo scontro con l'Anm prima del referendum"'. 11 April. Available at: https://www.repubblica.it/ (Accessed: 11 October 2026).

Linux Professional Institute (LPI) (n.d.). Learning Materials: Time Synchronization. Available at: https://learning.lpi.org/en/learning-materials/102-500/108/108.1/108.1_02/.

Microsoft (n.d.). Time Synchronization May Not Succeed on Windows Server. Available at: https://learn.microsoft.com/en-us/troubleshoot/windows-server/active-directory/time-synchronization-not-succeed-non-ntp.

National Security Agency (NSA) (n.d.). Official website. Available at: https://www.nsa.gov/.

NordVPN (n.d.). What Is NAT Firewall? Available at: https://nordvpn.com/blog/what-is-nat-firewall/.

Oxylabs (n.d.). Official website. Available at: https://oxylabs.io.

ProxyRack (n.d.). Official website. Available at: https://proxyrack.com.

Rid, T. and Buchanan, B. (2015). 'Attributing Cyber Attacks'. Journal of Strategic Studies, 38(1–2), pp. 4–37. https://doi.org/10.1080/01402390.2014.977382.

Schmitt, M.N. (ed.) (2017). Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations. 2nd edn. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316822524.

Smartproxy (n.d.). Official website. Available at: https://smartproxy.com.

Snowden, E. (2019). Permanent Record. New York, NY: Metropolitan Books.

Società Italiana di Intelligence (SOCINT) (2024). Associazione Culturale e Scientifica per la Promozione della Cultura dell'Intelligence. Available at: https://www.socint.org/ (Accessed: 11 October 2026).

Storm Proxies (n.d.). Official website. Available at: https://stormproxies.com.

The Guardian (2013). Tor Stinks: NSA Presentation Document. Available at: https://www.theguardian.com/world/interactive/2013/oct/04/tor-stinks-nsa-presentation-document.

The Tor Project (2026). Tor Browser Manual and Relay Infrastructure Overview. Available at: https://www.torproject.org/(Accessed: 11 October 2026).
The Wall Street Journal (2026). 'Boeing Is Pushing to Withdraw Guilty Plea Agreement'. Available at: https://www.wsj.com/ (Accessed: 11 October 2026).
Washington University in St. Louis (n.d.). The Power of Virtual Private Networks (VPN) in Privacy Protection. Available at: https://informationsecurity.wustl.edu/the-power-of-virtual-private-networks-vpn-in-privacy-protection/.
Wikipedia (n.d.). Cyberactivism. Available at: https://en.wikipedia.org/wiki/Hacktivism.
Wikipedia (n.d.). Denial-of-Service Attack. Available at: https://en.wikipedia.org/wiki/Denial-of-service_attack.
Wikipedia (n.d.). Edward Snowden. Available at: https://en.wikipedia.org/wiki/Edward_Snowden.
Wikipedia (n.d.). IPv6. Available at: https://en.wikipedia.org/wiki/IPv6.
Wikipedia (n.d.). Man-in-the-Middle Attack. Available at: https://en.wikipedia.org/wiki/Man-in-the-middle_attack.
Wikipedia (n.d.). Network Address Translation. Available at: https://en.wikipedia.org/wiki/Network_address_translation.
Wikipedia (n.d.). Overlay Network. Available at: https://en.wikipedia.org/wiki/Overlay_network.
Wikipedia (n.d.). Proxy Server. Available at: https://en.wikipedia.org/wiki/Proxy_server.
Wikipedia (n.d.). The Tor Network. Available at: https://en.wikipedia.org/wiki/Tor_(network).
Wikipedia (n.d.). Virtual Private Network. Available at: https://en.wikipedia.org/wiki/Virtual_private_network.

TURNING DIGITAL TRACES INTO RELIABLE EVIDENCE

Paolo Dal Checco
Forenser Srl, Italy

Matteo Vinci
Forenser Srl, Italy

Marco Musumeci
External Consultant for Forenser Srl, Italy

Abstract
Digital evidence typically begins as routine operational logs rather than forensic data, requiring digital forensics to transform these everyday records into legally viable proof. Written for non-specialists in cybersecurity, law, and compliance, this text outlines how the shift from technical traces to verifiable evidence intersects with key European frameworks, including the NIS2 Directive (incident reporting and chain-of-custody), the Cyber Resilience Act (product and vulnerability tracking), the AI Act (investigating AI behavior), and the Cross-Border Evidence Framework (enabling 8-hour emergency data orders). Crucially, the text emphasizes that regulation cannot override technical reality, meaning data that was never generated or has been overwritten is permanently lost. Consequently, organizations must embed digital forensics into their daily discipline by knowing where data is generated, documenting first-response actions, and keeping verifiable records.

1. From Digital Traces to Legal Proof

Many digital investigations begin with traces that were created for routine operations, not for a court file. An authentication timestamp, a cloud event, a help-desk ticket or a software update record can become important only later, when someone has to reconstruct what happened. At that point, it is not enough to ask what the data appear to show. The relevant questions are also where the data came from, whether they have been preserved properly and whether the surrounding steps were documented.

Digital Forensics manages that passage from technical trace to evidence. It does not make every trace probative; it asks whether the trace was collected lawfully, preserved with the least necessary alteration and interpreted in the right context. The same approach may serve an internal investigation, a regulatory report, a criminal proceeding or a civil dispute. In all of these settings, digital material can lose value quickly.

Why This Matters

The basic point for a non-specialist is that evidence is often lost before anyone recognises it as evidence. The first hours of an incident, the way logs are configured and the ordinary habits of documentation could determine whether the facts can be reconstructed later.

Forensic readiness should therefore be treated as preparation, not as a separate activity that starts only after a failure. Organisations do not need every employee to become a forensic examiner. They do need to know where useful records are created, how long they are kept and who can preserve them quickly.

Forensic Readiness in Practice

Forensic readiness means planning for evidence before the specific dispute or incident is known. A useful first step could be an evidence map: a plain inventory of systems, providers and workflows that may contain relevant traces. It should include obvious sources such as authentication logs, laptop and server records, firewall events, cloud audit trails and AI query logs. It should also include sources that are easy to forget: provider notifications, backup records, vulnerability tickets, change approvals and the emails exchanged during the first response.

In practice, cases cluster into a few recurring patterns: acquiring and analysing a device, account or cloud service; suspected data misuse by an employee; disputes over security measures; cyberattacks and fraud, for example business email compromise or SIM-swap; and checks on the authenticity of chats, images, audio or video.

A cloud service that exposes only recent management events may be sufficient for administration but inadequate for a later reconstruction. AWS CloudTrail event history, for example, covers the past 90 days of management events unless a trail or event data store has been configured for longer retention. The limitation is not simply a tool problem, it shows that retention is a prior organisational choice.

Readiness also affects the first response. Technical teams may have to isolate a server, reset credentials, restore a backup, export a virtual machine or disable an account. Waiting for a full forensic examination may be unrealistic. These actions are often necessary, but they are not neutral: they can change timestamps, overwrite logs, alter system states or remove material that would help explain the incident.

The practical answer is to make the response traceable. Even a short note stating who acted, what was done, when and why can be decisive later. It helps separate traces left by the attacker from traces created by the organisation's own response.

Table 1 Practical evidence map for non-specialists

Source	Why it matters	Practical check
Access and identity logs	Can show account use, failed logins, privilege changes or unusual locations.	Know the retention period and export method.
Cloud audit records	Can show console actions, API calls, resource creation or deletion.	Check whether logs cover management events only or also data events.
Backups and snapshots	Can preserve earlier states of files, systems or virtual machines.	Document when restoration occurs and what it overwrites.
Help-desk and change tickets	Explain why an action was taken and who authorised it.	Keep tickets linked to technical timestamps.
Vulnerability and update history	Can show whether a known weakness, patch delay or unsupported version was relevant.	Record versions, update dates and vendor notices.

2. European Rules and Forensic Consequences

European regulation does not create a single law of Digital Forensics. It does, however, affect the practical conditions in which evidence is generated, kept and requested. The method itself is older, resting on ISO/IEC 27037 and the Budapest Convention on Cybercrime.

For this chapter, four instruments are especially relevant:

- NIS2 addresses cybersecurity governance and incident reporting;
- the CRA adds the product perspective, focusing on products with digital elements;
- the AI Act regulates artificial intelligence through a risk-based model;
- the e-Evidence package deals with cross-border access to electronic evidence.

They do not replace forensic method. They shape the environment in which forensic method is applied.

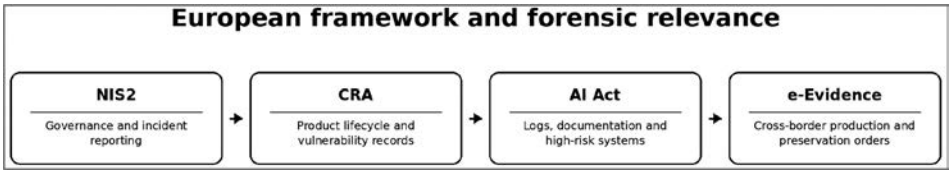

Fig. 1: European framework and forensic relevance

Under NIS2, cybersecurity is no longer treated as a purely technical matter: management bodies are required to approve and oversee cybersecurity risk-management measures. For Digital Forensics, this has a concrete consequence: poor logging, unclear procedures or undocumented first-response actions may point not only to a technical gap, but also to weak governance. NIS2 also gives CSIRTs a role that touches forensic work, including the collection and analysis of forensic data for risk and incident analysis. ENISA had already highlighted that first responders and CSIRTs may need practical guidance when handling electronic evidence. Incident reporting and forensic preservation remain distinct tasks, but they have to work together.

The CRA brings the product side into view. It sets horizontal requirements for products with digital elements; reporting obligations begin in September 2026 and the main obligations apply from December 2027. For an investigator, the practical value lies in the records surrounding the product: technical documentation, security-update information, support periods and vulnerability-handling history. These materials will not guarantee a complete reconstruction, but they can help identify whether an incident involved a known weakness, a missed update, a design or maintenance failure, a vulnerability or an attack technique not previously known.

The AI Act classifies AI systems by risk and imposes stricter duties on high-risk systems, especially where safety or fundamental rights may be affected. Requirements such as technical documentation, record-keeping, human oversight, and cybersecurity may become relevant when a disputed outcome has to be reconstructed. The investigator must first understand whether the AI system was only a tool used to process information, or whether it was itself part of the event under review. In the latter case, logs and documentation help to understand how the system was intended to operate and how it actually behaved.

The e-Evidence package deals with a familiar problem: relevant electronic evidence may be held by a provider in another Member State. From August 2026, competent authorities can issue European Production Orders and European Preservation Orders directly to service providers in other Member States. The Commission describes response times of 10 days in ordinary cases and 8 hours in emergencies. Faster access matters, especially for volatile data. It is still only useful if the request is proportionate, data protection is respected and the path by which the evidence was obtained is recorded clearly.

3. Operational Limits and Attribution

Technical limits remain even where the legal framework is clear: encryption can leave a device or file permanently unreadable if the PIN or password is never provided, and no vulnerability allows a way around it. On many smartphones this also depends on whether the device has already been unlocked at least once

since it was last switched on: file-based encryption leaves more data readable in that state, and a restart returns everything to a protected, largely unreadable condition. A message may have disappeared because the service was designed to make it disappear. Some systems also allow a duress PIN or password to delete data when entered. Standard cloud histories may be short unless longer retention was configured in advance.

Attribution has to be handled with the same caution. A device identifier or a login event can be important, but none of them is a "smoking gun" by itself. An IP address may point to a home connection, a corporate proxy, a compromised server or a VPN. A login event may show that credentials were used, but not who was at the keyboard.

Synthetic media make this harder today: a voice note, image or video, including CCTV or drone footage, cannot be treated as authentic only because it looks or sounds convincing. Technical analysis can examine metadata, compression patterns, editing traces and internal inconsistencies, but context remains essential: where the file came from, how it was shared, whether earlier versions exist and whether independent evidence points in the same direction. Digital Forensics is strongest when technical findings are supported by context, not presented as isolated conclusions.

Practical Steps: Preparing Evidence Before it is Needed

The points below are not a full forensic protocol and do not replace specialist procedures. They are meant as practical steps for non-specialists who may be involved in the first stages of an incident, when evidence is easiest to lose or confuse.

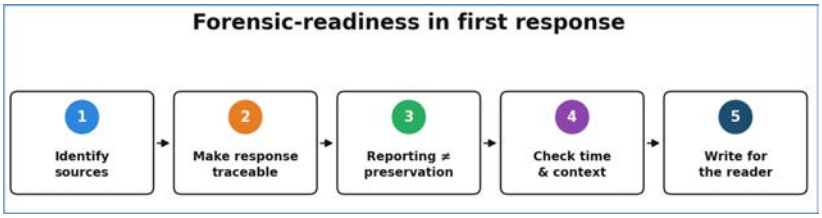

Fig. 2: Five evidence-aware practices for incident response

- Identify evidence sources in advance: list the systems that create relevant traces – identity platforms, endpoint tools, cloud consoles, backups, tickets, email alerts and vulnerability records. For each source, add the retention period and the person or team able to export it.
- Make first response traceable: containment and recovery may be urgent, but they should not be invisible. Record isolation, shutdowns and significant configuration changes as they happen. When data are copied, calculate hash values so that later readers can check that the copy has not changed.
- Keep reporting and preservation distinct: a regulatory notification de-

scribes an incident; preservation protects the material needed to verify it later. The two activities should inform each other, but they are not the same.
- Check time and context: record time zones, clock drift, log sources and the assumptions used to correlate events.
- Write for the later reader: a technical report should be readable by managers, lawyers and judges without hiding its technical basis. State what the data show, what they do not show and which alternative explanations were considered.

Practical Note: Preserve Before Interpreting

Before drawing conclusions, preserve volatile sources that may disappear: active sessions, short-retention logs, cloud events, chat records and the notes of the people who first handled the incident. Incident responders often have to work fast, but acting based on an early interpretation can damage the traces that could be needed later on.

In practice, any action that changes the state of a system should be recorded before or immediately after it is taken. The note does not replace specialist forensic work. It gives the specialist a record that can be checked.

4. Conclusion

Digital Forensics does not guarantee that every incident can be fully reconstructed. Its value is more limited, but still important: it reduces the risk that relevant traces are lost before their importance is understood. The European regulatory framework, including the NIS2 Directive, the Cyber Resilience Act, the AI Act and the e-Evidence package, does not remove technical limits such as inaccessible smartphones, disappearing messages or duress codes. It does, however, make governance, documentation and preservation more important.

The practical lesson of this chapter is forensic readiness: organisations must prepare for evidence before they need it. For companies, public bodies and judicial authorities, this means mapping data sources, documenting first-response actions as they happen, and keeping records that can be understood and verified. Where traces are fragile and retention periods are often short, this discipline can make the difference between a reconstruction that can be supported and a suspicion that remains unproven.

References

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 (NIS2 Directive), https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng

Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act), https://eur-lex.europa.eu/eli/reg/2024/2847/oj/eng

European Commission, Cyber Resilience Act, https://digital-strategy.ec.europa.eu/en/policies/cyber-resilience-act

European Commission, Cyber Resilience Act – Reporting obligations, https://digital-strategy.ec.europa.eu/en/policies/cra-reporting

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

European Commission, AI Act, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Regulation (EU) 2023/1543 of the European Parliament and of the Council of 12 July 2023 on European Production Orders and European Preservation Orders for electronic evidence in criminal proceedings and for the execution of custodial sentences following criminal proceedings, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32023R1543

Directive (EU) 2023/1544 of the European Parliament and of the Council of 12 July 2023 laying down harmonised rules on the designation of designated establishments and the appointment of legal representatives for the purpose of gathering electronic evidence in criminal proceedings, https://eur-lex.europa.eu/eli/dir/2023/1544/oj/eng

European Commission, E-evidence – cross-border access to electronic evidence, https://commission.europa.eu/law/cross-border-cases/judicial-cooperation/types-judicial-cooperation/e-evidence-cross-border-access-electronic-evidence_en

ENISA, Electronic evidence – a basic guide for First Responders, 2015, https://www.enisa.europa.eu/publications/electronic-evidence-a-basic-guide-for-first-responders

NIST, Special Publication 800-86, Guide to Integrating Forensic Techniques into Incident Response, 2006, https://csrc.nist.gov/pubs/sp/800/86/final

ISO/IEC 27037:2012, Information technology – Security techniques – Guidelines for identification, collection, acquisition and preservation of digital evidence, https://www.iso.org/standard/44381.html

Council of Europe, Convention on Cybercrime (Budapest Convention), CETS No. 185, 2001, https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=185

AWS Documentation, Working with CloudTrail event history, https://docs.aws.amazon.com/awscloudtrail/latest/userguide/view-cloudtrail-events.html

GrapheneOS, Features overview – Duress PIN/Password, https://grapheneos.org/features

Signal Support, Set and manage disappearing messages, https://support.signal.org/hc/en-us/articles/360007320771-Set-and-manage-disappearing-messages

FROM CVES TO CYBERSPACE SITUATIONAL UNDERSTANDING: RETHINKING VULNERABILITY INTELLIGENCE IN THE AGE OF ARTIFICIAL INTELLIGENCE

Andrea Accetta
NATO, Belgium

Abstract
The rapidly increasing volume, velocity, and complexity of software vulnerabilities are exceeding the capacity of traditional, human-centric vulnerability management processes. Since its inception, the Common Vulnerabilities and Exposures (CVE) framework has provided a standardized foundation for identifying and cataloguing individual software flaws. However, its inherently list-based approach is increasingly insufficient for representing the systemic and interconnected nature of cyber risk across today's complex digital ecosystems. This chapter explores how recent advances in artificial intelligence (AI), particularly systems capable of advanced reasoning, large-scale correlation, and predictive analysis, are transforming vulnerability intelligence from a reactive, disclosure-driven activity into a continuous and anticipatory capability. From 2019 to 2025, published CVEs expanded by almost 180%, representing a compound annual growth rate of 18.7%. Drawing on recent trends in CVE publication volumes and the widening gap between disclosed vulnerabilities and analysts' capacity to assess them, the chapter argues for a paradigm shift from managing discrete CVEs to developing comprehensive cyberspace situational understanding. In this emerging paradigm, AI-augmented vulnerability triage, correlation, and threat modelling become essential enablers of organisational and national cyber resilience. The chapter further examines the operational, governance, and ethical implications of delegating elements of the vulnerability management lifecycle to AI-enabled systems, including automation bias, adversarial manipulation of AI-assisted assessments, and the continuing necessity of meaningful human oversight. It concludes by proposing a forward-looking framework for integrating trustworthy AI capabilities into European and Allied vulnerability intelligence processes.

1. Introduction

For more than two decades, the Common Vulnerabilities and Exposures (CVE) programme has served as the lingua franca of vulnerability management. A CVE identifier allows a vendor in Luxembourg, a security operations centre in Brussels, and a national agency in Rome to refer unambiguously to the same flaw, and the enrichment layered on top of it (severity scores, affected products, weakness categories) has become the scaffolding around which scanners, compliance regimes, and patching policies have been built. The significance

of this achievement is considerable: a shared vocabulary is a precondition for collective defence.

A vocabulary, however, is not an understanding. A list of identifiers, however well curated, tells a decision-maker what is broken; it does not tell them what matters, what will be attacked, or what the consequences of an attack would be for the missions that depend on the affected systems. Closing that gap between the catalogue and the decision has always required human analysis: someone had to read the entry, relate it to the organization's systems, and judge its importance. Two developments have now broken that arrangement. The number of published vulnerabilities has grown faster than the human capacity to assess them, so that the analytical step is being skipped or performed too late; and artificial intelligence (AI) has begun to discover and exploit flaws at a speed no human research team can match, so that the interval in which analysis is useful has shrunk. The result is that the gap between cataloguing vulnerabilities and understanding cyber risk has become the central problem of vulnerability intelligence, and AI is at once the reason the problem has become acute and the means by which it may be addressed.

This chapter makes a single argument with two parts. The objective of vulnerability intelligence should change, from managing discrete CVEs to building and maintaining cyberspace situational understanding; and AI is the means by which that objective becomes attainable at the scale and tempo the problem now demands, provided it is governed so that human judgement remains at the point of decision. The answer to the problem, in other words, is not a faster list. It is a different objective, and a different way of working. The perspective adopted is deliberately operational: it draws on the author's experience of building and sustaining cyberspace situational awareness in a multinational defence environment, and it asks not only what vulnerability intelligence should become but how the picture is actually built, maintained, and used to decide, day after day. Vulnerability intelligence[1] should be re-oriented away from the management of discrete CVEs and towards the building and maintenance of *cyberspace situational understanding*: a continuously updated, mission-referenced picture of where an organization, a Member State, or the Union as a whole is exposed, how that exposure is likely to be exploited, and what should be done about it. The chapter is presented in seven further sections. Section 2 documents the scale of the vulnerability deluge. Section 3 distinguishes situational awareness

[1] Vulnerability intelligence is understood here as the systematic collection, contextualization, and prioritization of information about vulnerabilities in support of risk-informed decisions. Industry usage treats it as a specific form of threat intelligence, distinguished from vulnerability management by its emphasis on exploitability, adversary activity, and prioritization rather than on scanning and patching (Flashpoint, 2026). The chapter adopts that usage and extends it, arguing that the discipline's proper output is situational understanding rather than a prioritized list.

from situational understanding, explains why the distinction matters for cyberspace, and sets out the European legal architecture within which vulnerability information has become an object of public governance. Section 4 examines the AI inflection point of 2024–26. Section 5 describes, from a practitioner's standpoint, how a situational picture is built and maintained and where AI changes the process, and Section 6 addresses the governance conditions under which AI can be trusted with part of that task. Section 7 translates the argument into a framework and a short operational playbook for European vulnerability intelligence, before a concluding reflection.

2. The Vulnerability Deluge

The quantitative trend is unambiguous. As Figure 1 shows, the number of CVEs published annually rose from roughly 17,300 in 2019 to more than 48,000 in 2025 (CVE Details, 2026), an increase of almost 180%, equivalent to a compound annual growth rate of 18.7% on the author's calculation. Growth accelerated markedly after 2023, and the trajectory has not flattened: submissions in the first quarter of 2026 were running about a third above the same period of the previous year, and some industry forecasts anticipate 70,000 or more CVEs for the year as a whole (Infosecurity Magazine, 2026).

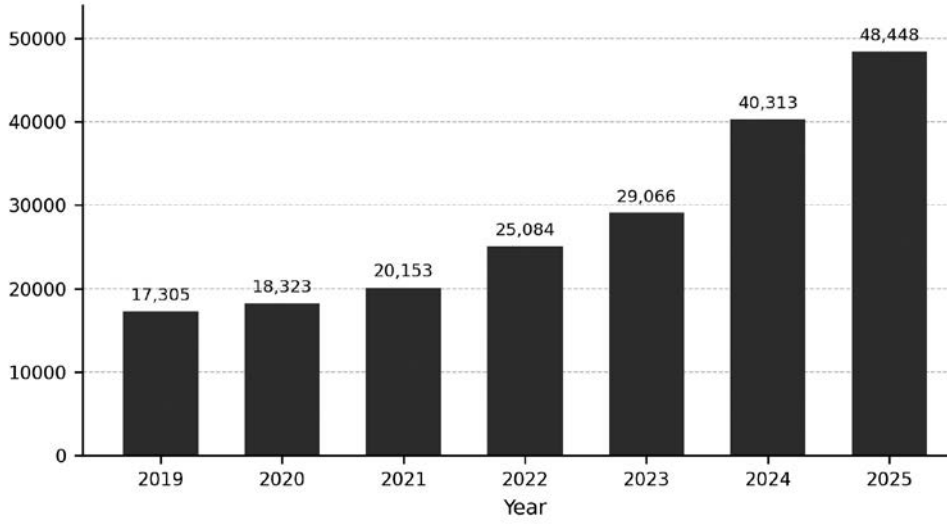

Fig. 1. Growth in the volume of published CVEs, 2019–2025 (source: author's elaboration of CVE Details, 2026; copyright: Andrea Accetta, 2026).

The consequences of this growth for the institutions that make CVE data

usable became visible in April 2026, when the United States National Institute of Standards and Technology (NIST) announced that its National Vulnerability Database (NVD) would no longer attempt to enrich every submitted CVE with severity scores, product mappings, and weakness classifications. NIST reported that submissions had grown by 263% between 2020 and 2025 and that, despite enriching a record 42,000 CVEs in 2025, it could not keep pace; enrichment would henceforth be prioritized for vulnerabilities under known exploitation, in federal software, or in software designated as critical, while a backlog of tens of thousands of older entries was moved to a "Not Scheduled" status (NIST, 2026). Industry analysts estimated that the prioritized categories would cover only 15–20% of anticipated CVE volume (Cloud Security Alliance, 2026).

The problem, then, is that the institution on which most of the world relies for usable vulnerability data can no longer process what it receives. The problem is not confined to the supply side of vulnerability data, however. On the demand side, organizations are falling further behind. The Verizon *2026 Data Breach Investigations Report*, drawing on its global dataset of confirmed breaches, found that exploitation of vulnerabilities had become the most common initial access vector for the first time in the report's history, accounting for 31% of breaches, up from 20% the previous year; at the same time, the median time to patch known-exploited vulnerabilities lengthened from 32 to 43 days, and only about a quarter of such vulnerabilities were fully remediated within the year (Verizon, 2026: [p.]). Mandiant's frontline data tell a complementary story: exploits remained the leading initial infection vector for the sixth consecutive year, and, across the vulnerabilities weaponized in 2025, exploitation on average preceded the public availability of a vendor patch (Mandiant, 2026).

Collectively, these figures describe a system in which more flaws are disclosed than can be enriched, more are enriched than can be triaged, and more are triaged than can be remediated before adversaries act. Prioritization mechanisms such as the Forum of Incident Response and Security Teams' (FIRST) Exploit Prediction Scoring System (EPSS) and the Cybersecurity and Infrastructure Security Agency's (CISA) Known Exploited Vulnerabilities (KEV) catalogue offer relief, and the empirical case for them is strong: only about 0.5% of published CVEs appear in KEV, and in any thirty-day window only 2–3% show observed exploitation activity (FIRST, 2026; Jacobs *et al.*, 2021). The implication of these base rates is central to this chapter's argument: the problem is not that 48,000 flaws must be fixed each year, but that the few hundred that will matter to a given organization must be found inside the 48,000, quickly, and in context. That is a task of selection under uncertainty at a scale that manual review cannot sustain, which is precisely why automated, context-aware prioritization has become a necessity rather than an optimization. But these mechanisms are, by design, global rather than local. They estimate

whether a flaw is likely to be exploited somewhere, not whether it is reachable, exposed, and mission-relevant in a specific network. That final, decisive step of contextualization remains overwhelmingly manual.

3. From Situational Awareness to Situational Understanding

The concept of situational awareness (SA) has a well-established theoretical basis. Endsley (1995) conceptualized SA as a three-level construct consisting of the perception of relevant elements in the environment, the comprehension of their meaning, and the projection of their status into the near future. Endsley's model was developed for physical, time-critical operational settings, aviation above all, and its transfer to cyberspace is not self-evident. It is adopted here for three reasons. First, it is explicitly decision-oriented: it defines awareness by what an operator must know in order to act under time pressure, which is the condition of a vulnerability analyst as much as of a pilot. Second, it is the model on which the cyber SA literature itself has largely been built: the foundational edited volume by Jajodia et al. (2010) frames cyber SA in Endsley's terms, and the systematic review by Franke and Brynielsson (2014) identifies it as the predominant theoretical reference in a field they otherwise found to be growing but fragmented and lacking an integrated body of theory. Third, the model's separation of perception from comprehension and projection is precisely the distinction that vulnerability management, as currently practised, fails to make. The transfer nonetheless requires care. Cyberspace differs from a cockpit in that the terrain is human-made and constantly modified, the sensors are themselves software subject to the same flaws they observe, and the adversary is an active, adaptive agent rather than weather. These differences do not invalidate the model; they explain why, in cyberspace, awareness in Endsley's sense is easier to obtain and less sufficient than elsewhere, and they motivate the chapter's emphasis on understanding. Mapped onto vulnerability intelligence, Level 1 corresponds to the perception of raw indicators: CVE feeds, scanner output, asset inventories, exploit telemetry. Level 2 corresponds to comprehension: knowing which of these indicators matter, in which context, and why. Level 3 corresponds to projection: anticipating which exposures will be exploited, along which paths, and with what consequence.

Most vulnerability management programmes today are strong at Level 1 and weak at Levels 2 and 3. They perceive a great deal and understand comparatively little. It is for this reason that this chapter deliberately speaks of situational understanding rather than awareness. The distinction should be stated precisely, because the remainder of the chapter relies on it. Understanding is not a fourth level added to Endsley's three; it is what the three functions

produce when they are performed continuously, referenced to a mission, and connected to a decision. Perception, comprehension, and projection are therefore treated here as the constituent functions of understanding rather than as a linear sequence through which data passes once, and the tables and figure that follow use them in that sense. In military doctrine, the distinction is not merely semantic. The United Kingdom's Joint Doctrine Publication 04 treats understanding as the interpretive layer that converts perception of a situation into the context, insight, and foresight on which effective decision-making depends (Ministry of Defence, 2016: [p.]). Awareness is knowing that something is happening; understanding is knowing what it means for one's own mission and what to do about it. In cyberspace, where the operating environment is a human-made, constantly modified, and largely opaque terrain, the difference between the two is the difference between a dashboard and a decision.

A second reason for the shift in terminology concerns the relationship between situational awareness and cyber risk. Situational understanding is not an end in itself; it is the input to risk assessment, and risk is a judgement about the likelihood that an exposure will be exploited and the consequence if it is. Both terms of that judgement require comprehension and projection, not perception alone. The objects of vulnerability intelligence, moreover, are themselves changing. A CVE describes a single flaw in a single product. Cyber risk, however, is systemic and relational: it emerges from the composition of components, from transitive dependencies in software supply chains, from the privilege relationships that allow an attacker to move from a low-value foothold to a domain controller, and from the mission dependencies that determine whether the loss of a given server is an inconvenience or a crisis. Verizon's analysis of attack graphs found that, in 16% of organizations studied, an attacker with low-level initial access had an 80% or better chance of compromising a key administrative asset (Verizon, 2026: [p.]). No CVE record captures that property, because it is a property of the system, not of the flaw.

Vulnerability Information as an Object of European Governance

The European Union has, over the past four years, constructed something that no other jurisdiction possesses: a coherent legal architecture in which vulnerability information is treated as an object of public governance rather than as a by-product of vendor disclosure. Four instruments are central. The NIS2 Directive requires Member States to establish coordinated vulnerability disclosure policies and tasks the European Union Agency for Cybersecurity (ENISA) with maintaining a European vulnerability database (European Union, 2022). That database, the European Vulnerability Database (EUVD), was declared operational in May 2025 while still described by ENISA as under development; it aggregates information from national Computer Security

Incident Response Teams (CSIRTs), vendors, and existing registries including the CVE programme, and presents dedicated views of critical vulnerabilities, exploited vulnerabilities, and vulnerabilities coordinated by European CSIRTs (European Commission, 2025; Help Net Security, 2025).

The Cyber Resilience Act goes further, imposing on manufacturers of products with digital elements a binding duty to report actively exploited vulnerabilities and severe incidents: an early warning within 24 hours, a full notification within 72 hours, and a final report within 14 days of a corrective measure, all filed through a Single Reporting Platform operated by ENISA and shared with the competent national CSIRT (European Union, 2024b; European Commission, 2026). The Cyber Solidarity Act adds the operational layer, establishing a European Cybersecurity Alert System of national and cross-border cyber hubs whose purpose is to detect and share information on threats across borders and, increasingly, to do so with the aid of advanced analytics and AI (European Union, 2025). The Artificial Intelligence Act, finally, requires high-risk AI systems to achieve appropriate levels of accuracy, robustness, and cybersecurity, and thereby establishes the legal vocabulary in which the trustworthiness of AI-enabled defensive tools will be assessed (European Union, 2024a).

Collectively, these instruments give the Union a distinctive starting position, but it is a position on paper whose operational value remains to be demonstrated. The reporting obligations of the Cyber Resilience Act apply only from 11 September 2026; the Single Reporting Platform through which they are discharged was scheduled to go live on the same date, and at the time of writing there is no experience of how manufacturers will comply, how national CSIRTs will absorb the volume, or how much of what is reported will be usable in practice. The EUVD, for its part, continues to depend on the CVE programme and other upstream sources whose own funding and operating models have proved fragile, as the events of 2025 and 2026 at MITRE and NIST demonstrated. What the architecture is designed to produce is a continuous, legally mandated stream of exploitation intelligence with no equivalent elsewhere; whether it does so, and with what latency and quality, will be known only after the first years of operation. Even if it functions as intended, a stream of reports is perception, not comprehension. The governance architecture addresses data provenance and availability; it does not, by itself, address understanding. The contribution of this chapter is to ask how the second problem can be addressed on the foundation the first has laid (in operational terms, how the reports become a picture and the picture becomes a decision) and how AI can be used to do so in a manner consistent with the values the same legal framework embodies.

4. The AI Inflection

Between 2024 and 2026, the relationship between AI and vulnerability discovery passed from the speculative to the demonstrated. Three developments are of particular relevance. First, academic work showed that large language model (LLM) agents, given a CVE description and appropriate tooling, could autonomously produce working exploits for a majority of a sample of real-world one-day vulnerabilities (Fang *et al.*, 2024). Second, in November 2024 Google's Project Zero and DeepMind reported that their Big Sleep agent had found a previously unknown, exploitable memory-safety bug in SQLite, described by the team as the first public example of an AI agent discovering such a flaw in widely used software, building on results from the Defense Advanced Research Projects Agency's AI Cyber Challenge (Google Project Zero and Google DeepMind, 2024). Third, in April 2026 Anthropic announced that its Claude Mythos Preview model could autonomously discover and exploit zero-day vulnerabilities across every major operating system and browser; rather than release the model publicly, the company deployed it through a controlled defensive coalition, Project Glasswing, whose partners reported more than 10,000 high- or critical-severity findings within roughly six weeks, including a 27-year-old flaw in OpenBSD and a certificate-forgery vulnerability in the wolfSSL cryptography library (Anthropic, 2026a; 2026b). The *International AI Safety Report 2026* concluded that vulnerability discovery is among the domains where evidence of meaningful AI uplift is strongest (Bengio *et al.*, 2026).

The implications for vulnerability intelligence are profound and double-edged. On the defensive side, the cost of finding flaws before adversaries do has fallen dramatically, and code that has resisted decades of human review is now being audited at machine scale. On the offensive side, the same capability compresses the window between disclosure and weaponization. Mandiant's observation that exploitation now, on average, precedes patch availability (Mandiant, 2026) predates the broad diffusion of frontier models; the plausible near-term future is one in which AI-discovered flaws, AI-generated exploits, and AI-orchestrated intrusions raise the tempo of the vulnerability lifecycle beyond anything human-centred processes were designed for. NIST's decision to abandon universal enrichment can be read, in this light, as the first institutional acknowledgement that the legacy pipeline cannot absorb AI-scale discovery (NIST, 2026).

Consider how this plays out operationally. A critical flaw in a widely deployed network appliance is disclosed late on a Friday. In the legacy process, the security team waits for enrichment, checks the scanner on Monday, raises a change request, and patches within the month; the adversary, equipped with an AI agent that generated a working exploit within hours of disclosure, has

been inside the network since Saturday. In an AI-augmented process, the disclosure is correlated within minutes against the terrain map: the appliance is identified as key terrain for two mission functions, its exposure to the internet is confirmed from configuration data, the absence of a compensating control is noted, and an exploit-path projection shows a route to the identity infrastructure. The duty officer receives one recommendation, with its reasoning, and authorizes an emergency mitigation on Friday night. The difference is not that the second organization has better data; it is that it has a picture, and a process for updating it, into which the data can fall.

This illustration is stipulative rather than evidential: it presumes an accurate terrain map, current configuration data, and a correlation step that produces no false positives, conditions that Section 5 shows to be demanding. The wider point it illustrates is that AI cannot be treated merely as one more tool inside the existing vulnerability management workflow. It is simultaneously the principal driver of the problem and, on the argument advanced here, the most promising route to a solution. The remainder of this chapter takes that dual character as its premise.

5. Building and Maintaining the Picture: an Operational Perspective

Situational understanding is not produced by a tool alone, nor, at the scale described in Section 2, by analysts alone; it is produced by a human–AI team working through a process in which each party does what the other cannot. In operational practice, that process has a recognizable shape regardless of the organization that runs it, and it should be described plainly, because the places where AI changes the process are only visible once the process itself is made explicit. Six elements recur, summarized in Table 1.

Mission Decomposition and Cyber Key Terrain

The process begins not with the vulnerability feed but with the mission. An operational headquarters, a critical infrastructure operator, or a Member State authority first decomposes what it must achieve into the functions on which that achievement depends, and then maps those functions onto the systems, services, data flows, and identities that deliver them. The result is a map of *cyber key terrain*: the assets whose compromise would degrade or deny the mission. This map is the reference against which every subsequent vulnerability is judged, and without it prioritization is meaningless. It is also the element most often missing in civilian vulnerability management programmes, which tend to inherit an asset inventory rather than derive one from the mission. Building and maintaining the terrain map is laborious, and AI-assisted discovery of de-

pendencies, from network telemetry, configuration data, and documentation, is among the most valuable and least glamorous applications of the technology.

Information Requirements

An operational picture is built around the questions the decision-maker must answer, formalized as priority information requirements. For vulnerability intelligence the questions are concrete: which of our key terrain is exposed to a flaw that is being exploited now; which exposure would give an adversary a path from the perimeter to the terrain; which of our suppliers' products has just entered a reporting stream; what has changed since the last picture. The Cyber Resilience Act's reporting regime, the EUVD's exploited-vulnerabilities view, and global priors such as the Exploit Prediction Scoring System and the Known Exploited Vulnerabilities catalogue are answers to the *global* form of these questions. Converting them into the *local* form (does this concern my terrain, my configuration, my exposure) is the analytical core of the process, and it is where AI-augmented correlation delivers the greatest relief.

The Collection Plan

Each information requirement is served by identified sources. In vulnerability intelligence the collection plan spans internal telemetry (scanners, endpoint agents, configuration management, identity systems), regulatory streams (CSIRT notifications, EUVD, manufacturer reports under the Cyber Resilience Act), commercial and open sources, and, increasingly, the organization's own AI-assisted discovery runs against its own code and systems. The discipline that operational intelligence brings is the explicit mapping of requirement to source, so that gaps are known rather than assumed. A collection plan also makes the intelligence pipeline auditable, which becomes essential once part of it is automated.

Correlation, Exposure Assessment, and Projection

Correlation binds disclosed flaws to assets, reachability, privilege relationships, and observed adversary activity; projection reasons forward over attack graphs to anticipate exploitation paths, time-to-exploit, and mission impact. These are the comprehension and projection functions of the picture, and the two places where machine assistance most changes what a small team can achieve. The human contribution is adjudication: resolving conflicts between sources and models, assessing consequence, and deciding on courses of action. It is also at this point that the base rates of Section 2 do their work: correlation is what reduces tens of thousands of disclosures to the handful of exposures that concern a particular organization.

Decision Cadence and the Picture Itself

Operational organizations maintain their picture through a battle rhythm: a scheduled cycle of updates, briefings, and decision points at which human decision-makers act. The legacy vulnerability management cycle of monthly patching and quarterly scanning was a rhythm of this kind, designed for an era in which the interval between disclosure and exploitation was measured in weeks. That assumption no longer holds; exploitation now, on average, precedes patch availability (Mandiant, 2026). Two layers must therefore be distinguished. The machine layer has no rhythm: AI agents ingest, correlate, and re-score exposures continuously as telemetry and reporting streams change. The human layer retains a cadence, but it becomes event-driven rather than calendar-driven: decision points are triggered by material changes in the exposure state of key terrain, and the scheduled battle rhythm is reserved for review, direction, and the validation of the picture rather than for its production. The picture itself should be expressed not as a list of CVEs but as the *exposure state* of each element of key terrain, with an explicit assessment of likelihood, consequence, and available courses of action. Table 1 summarizes how the elements of the process contribute to understanding and how the work is divided between analyst and machine.

Table 1: Operational elements of vulnerability intelligence, their contribution to situational understanding, and the role of the human–AI team

Operational element	Function of understanding	Human role	AI contribution
Mission decomposition and cyber key terrain	Foundation	Define mission functions and decision priorities; validate terrain map	Infer dependencies from telemetry, configuration, and documentation; keep the map current
Information requirements	Foundation	Formulate the questions the decision-maker must answer	Translate global signals (EUVD, KEV, EPSS, CRA reports) into local questions
Collection plan	Perception	Map requirements to sources; own gaps	Ingest, normalize, and de-duplicate streams; run own-code discovery and variant analysis
Correlation and exposure assessment	Comprehension	Adjudicate conflicts; assess consequence	Correlate flaws with assets, reachability, privileges, and adversary activity

Operational element	Function of understanding	Human role	AI contribution
Exploit-path and impact projection	Projection	Decide courses of action; accept residual risk	Reason over attack graphs; forecast time-to-exploit; propose mitigations
Decision cadence and after-action review	Sustainment	Set event-driven decision points; review and direct; learn from outcomes	Operate continuously; surface material change; feed outcomes back to models

Exercises, After-action Review, and the Changing Role of the Analyst

Finally, operational organizations validate their picture by exercising it. Structured exercises reveal whether the terrain map is accurate, whether the information requirements are the right ones, and whether the decision points are placed where decisions can still be taken. The after-action review that follows is the mechanism by which the process improves, and, in an AI-augmented pipeline, the mechanism by which models are corrected against what actually happened rather than what was disclosed. In this setting the human role changes. When AI systems perform perception and much of comprehension, the analyst's scarce judgement is best spent on adjudication: validating machine-generated hypotheses, resolving conflicts between models and sources, deciding on consequence, and owning the decision. This is a more demanding role, not a diminished one, and it has implications for training, career structures, and the design of the human–machine interface that European organizations have only begun to confront.

6. Governance, Trust, and Meaningful Human Oversight

Delegating elements of the vulnerability lifecycle to AI introduces risks that are as real as the benefits. Three deserve particular attention.

The first is *automation bias*: the well-documented tendency of human operators to over-trust automated recommendations, to under-monitor automated processes, and to fail to intervene when automation errs (Parasuraman and Manzey, 2010). In vulnerability intelligence, automation bias manifests as the uncritical acceptance of machine-assigned priorities and, more insidiously, as the gradual atrophy of the analytical skills needed to challenge them. The proposed remedy is not less automation but better-designed oversight: calibrated confidence reporting, explicit uncertainty, deliberate red-teaming of model outputs, and periodic exercises in which analysts must reconstruct a machine's reasoning without its help.

The second is *adversarial manipulation*. A prioritization model is an attack surface. An adversary who understands which signals drive an organization's triage can flood those signals with noise, suppress indicators that would raise a target's priority, or seed poisoned data into the open sources on which enrichment depends. The reconciliation of divergent severity scores is already a contested space: analyses have found disagreement between NVD and vendor-supplied scores in more than half of dual-scored CVEs (Recorded Future, 2026); AI-mediated assessment is likely to make it more so. Provenance tracking, source diversity, anomaly detection on the intelligence pipeline itself, and adversarial testing of models against manipulation must be treated as core requirements rather than optional hardening.

The third is *accountability*. The Artificial Intelligence Act requires that high-risk AI systems be designed for effective human oversight, be sufficiently transparent for deployers to interpret their output, and maintain logs that allow their operation to be traced; it also requires resilience against attempts to manipulate the system's behaviour, including data poisoning and adversarial inputs (European Union, 2024a). Whether or not a given vulnerability intelligence application falls within the Act's high-risk categories, these requirements articulate the standard that a defensive AI system should meet. A prioritization system that cannot explain why an exposure was ranked as it was, that cannot be audited after a failure, or that cannot be disengaged when it behaves unexpectedly, does not meet it. Meaningful human oversight, in this reading, is not a slogan appended to an automated pipeline; it is an architectural property that must be designed in from the start. It is also, in a Union of sovereign Member States with differing legal traditions and levels of AI maturity, a precondition for the mutual trust on which shared situational understanding across borders depends. A further consideration follows from the framework itself and should be stated even though it cannot be fully resolved here. If comprehension and projection are increasingly performed by AI, then a decision taken on the basis of the picture is, in part, a decision taken on the basis of an AI projection, and the decision-maker's understanding of the situation is only as good as their understanding of the system that produced it. Situational understanding must therefore include a second-order component: an explicit, current account of the AI elements of the pipeline, covering which models are in use and in which version, what data they were last updated with, how their confidence is calibrated, where their known failure modes lie, and which parts of the picture are machine-generated rather than observed. In operational terms, the state of the AI should be an element of the picture in its own right, reported alongside the exposure state of key terrain, so that the line between what the organization knows and what its models infer remains visible at the point of decision. The principles at stake (lawfulness, accountability, explainability, reliability, govern-

ability, and bias mitigation) are also those on which the Union and its international partners have converged, which is itself a resource for cooperation.

7. Towards a Framework for European Vulnerability Intelligence

Figure 2 summarizes the argument of this chapter as a framework for integrating trustworthy AI into vulnerability intelligence. The left-hand column shows the three constituent functions of understanding, re-specified for cyberspace: perception of vulnerability and asset data; comprehension through contextualized vulnerability intelligence; and projection of exploitation paths and mission impact. They are drawn as a stack rather than a sequence: understanding, at the top, is the state that results when all three are sustained together and connected to decision and action, and the feedback path returns outcomes to the data layer so that the cycle is continuous rather than linear. The right-hand column identifies where AI augmentation applies (autonomous discovery and variant analysis, large-scale correlation and enrichment, predictive triage and exploit-path reasoning) and anchors all of it in meaningful human oversight, assurance, and accountability. The feedback arrow signifies that decisions and outcomes must flow back into the data layer so that models learn from what actually happened, not only from what was disclosed.

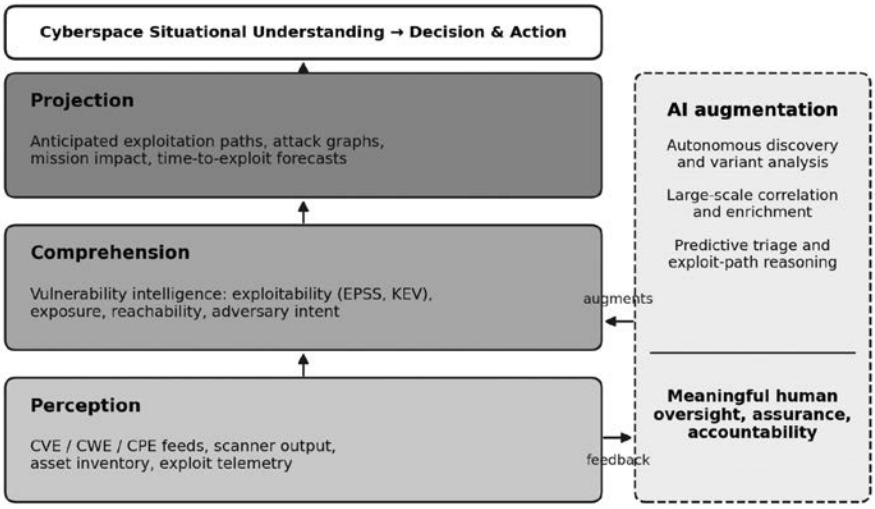

Fig. 2: From CVEs to cyberspace situational understanding: an AI-augmented framework (copyright: Andrea Accetta, 2026).

The framework maps directly onto the European architecture described in Section 3. The perception layer is, in the European case, defined by law even where it is not yet mature in practice: the EUVD, the Cyber Resilience Act reporting stream, and the notifications flowing through national CSIRTs and the European Cybersecurity Alert System are intended to constitute a data foundation that other regions must assemble voluntarily, and the early years of the Cyber Resilience Act reporting regime will show how far that intention is realized. The comprehension and projection layers are where the Union's investment should now concentrate, and where the framework's AI components apply. Four practical priorities follow.

First, a shared, mission-referenced exposure model, rather than a shared vulnerability list, should become the object of federated exchange between Member States, through the CSIRTs Network and the cross-border cyber hubs, so that situational understanding is built collectively while sovereignty over sensitive asset data is preserved.

Second, access to frontier vulnerability-discovery capabilities, currently concentrated in a small number of firms and controlled coalitions, should be secured for defensive use on European critical systems under agreed assurance conditions; the European Cybersecurity Competence Centre and the Digital Europe and Horizon Europe programmes offer the funding and coordination instruments through which such access, and the European capabilities that should eventually complement it, can be developed.

Third, common European standards for the provenance, explainability, and adversarial testing of AI-generated vulnerability assessments should be elaborated, building on the requirements of the Artificial Intelligence Act and on ENISA's certification role.

Fourth, the analyst workforce should be deliberately re-skilled for adjudication rather than processing, with exercises that keep human judgement sharp precisely where automation is strongest; the Cybersecurity Skills Academy provides a natural vehicle.

These priorities can be made concrete. Table 2 sets out, in the form of a short playbook, the actions that an operational organization (a Member State authority, a cross-border cyber hub, a critical infrastructure operator) can take to move from CVE management to situational understanding, with phases aligned to the 24-hour, 72-hour, and 14-day intervals that the Cyber Resilience Act imposes on manufacturers' reporting, adopted here as a convenient common tempo for defensive action. None of these priorities is exclusively European. They are shared with the Union's partners, including NATO, and the established channels of EU–NATO cooperation in cyber defence offer a means of pursuing them without duplication (European Union and NATO, 2023).

Table 2: An operational playbook: from CVE management to cyberspace situational understanding

Phase	Actions	Outcome
Establish (once, then maintain)	Decompose missions into functions; map cyber key terrain; formulate information requirements; write the collection plan; define exposure states and decision points; nominate the adjudicating analyst role	A mission-referenced picture against which every vulnerability can be judged
Continuous (machine tempo)	Ingest EUVD, CSIRT, CRA, KEV, EPSS, and internal telemetry; run own-code discovery and variant analysis; correlate to terrain; re-score exposures; log provenance and model reasoning	A picture that is current at all times and auditable after the fact
First 24 hours (on material change)	Confirm exposure of key terrain; project exploit paths; identify compensating controls; brief the decision-maker with one recommendation and its reasoning; where the organization is a manufacturer, file the CRA early warning	A decision taken while it still matters
72 hours	Execute and verify mitigation; extend correlation to suppliers and dependencies; share the exposure assessment with the CSIRT and, where relevant, the cross-border cyber hub	Contained exposure; shared understanding across borders
14 days and after	Conduct after-action review; correct terrain map, requirements, and models; record lessons; feed outcomes back into training data and exercises	A process that improves with every event

8. Conclusion: Towards a Terrain that Can Be Read

This chapter has argued that vulnerability intelligence should be reconceived as the discipline through which cyberspace situational understanding is built and maintained, rather than as the administration of a growing list of identifiers, and it has described that discipline in the operational terms of mission, terrain, requirements, collection, decision cadence, and review in which it is actually practised. The argument can be placed in a longer perspective. Successive transformations in military affairs have been, in large part, transformations in understanding: cartography, optics, radar, and satellite reconnaissance each extended a commander's capacity to know the terrain before committing forces

to it. Cyberspace has been an anomaly in this respect. It is the only operational domain that is entirely human-made, yet defenders have consistently known it less well than the adversaries who operate within it, and have been constrained to read their own terrain one vulnerability at a time.

The evidence reviewed in Section 4 suggests that this asymmetry may now be narrowing. For the first time, the terrain of cyberspace can be examined at the speed and scale at which it changes: AI systems can audit, within hours, code that decades of human review left unexamined, correlate exposures across millions of assets, and reason forward along the paths an intruder would be likely to take. The capabilities most likely to discover the next critical flaw are, moreover, available to defenders as well as to adversaries. Vulnerability intelligence, on this trajectory, ceases to be an inventory of known defects and approaches a continuously updated model of the operational environment: a cyberspace common operating picture that is perceived, comprehended, and projected without interruption.

The question confronting Europe is therefore not whether AI should be entrusted with elements of vulnerability intelligence, but how the Union, its Member States, and their partners should organize themselves so that such trust is warranted. Situational understanding is not a product to be procured; it is a discipline, sustained by institutions, standards, and analysts who know when to doubt their instruments. If that discipline can be established around the new capabilities, retaining human judgement at the point of decision while allowing machines to bear the burden of perception and comprehension, then Europe, which has done much to make vulnerability information a matter of public governance, may secure the advantage that terrain has always conferred on those who understand it best.

References

Anthropic. 2026a. *Assessing Claude Mythos Preview's Cybersecurity Capabilities*. Available online at https://red.anthropic.com/2026/mythos-preview (Last accessed: September 6, 2026).

Anthropic. 2026b. *Project Glasswing: An Initial Update*. Available online at https://www.anthropic.com/research/glasswing-initial-update (Last accessed: September 6, 2026).

Bengio, Y. *et al.* 2026. *International AI Safety Report 2026*. Available online at https://arxiv.org/abs/2602.21012 (Last accessed: September 6, 2026).

Cloud Security Alliance. 2026. *NVD Enrichment Triage: Enterprise Vulnerability Programs Must Adapt*. Research note. Available online at https://labs.cloudsecurityalliance.org/research/csa-research-note-nist-nvd-enrichment-policy-change-20260419/ (Last accessed: September 6, 2026).

CVE Details. 2026. *Vulnerabilities by Year*. Available online at https://www.cvedetails.com/browse-by-date.php (Last accessed: September 10, 2026).

Endsley, M.R. 1995. Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors* 37 (1): 32–64.

European Commission. 2025. *EU Launches a European Vulnerability Database to Boost its Digital Security*. Available online at https://digital-strategy.ec.europa.eu/en/news/eu-launches-european-vulnerability-database-boost-its-digital-security (Last accessed: September 6, 2026).

European Commission. 2026. *Cyber Resilience Act: Reporting Obligations*. Available online at https://digital-strategy.ec.europa.eu/en/policies/cra-reporting (Last accessed: September 6, 2026).

European Union. 2022. 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 (NIS 2 Directive). *Official Journal of the European Union* L 333: 80–152.

European Union. 2024a. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union* L, 12 July 2024.

European Union. 2024b. Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act). *Official Journal of the European Union* L, 20 November 2024.

European Union. 2025. Regulation (EU) 2025/38 of the European Parliament and of the Council of 19 December 2024 laying down measures to strengthen solidarity and capacities in the Union to detect, prepare for and respond to cybersecurity threats and incidents (Cyber Solidarity Act). *Official Journal of the European Union* L, 15 January 2025.

European Union and NATO. 2023. *Joint Declaration on EU–NATO Cooperation*, 10 January 2023. Brussels: European Union and North Atlantic Treaty Organization.

Fang, R., Bindu, R., Gupta, A. and Kang, D. 2024. *LLM Agents Can Autonomously Exploit One-Day Vulnerabilities*. Preprint. Available online at https://arxiv.org/abs/2404.08144 (Last accessed: September 6, 2026).

Flashpoint. 2026. *Vulnerability Intelligence*. Available online at https://flashpoint.io/intelligence-101/vulnerability-intelligence/ (Last accessed: September 10, 2026).

FIRST. 2026. *EPSS Frequently Asked Questions*. Forum of Incident Response and Security Teams. Available online at https://www.first.org/epss/faq (Last accessed: September 6, 2026).

Franke, U. and Brynielsson, J. 2014. Cyber Situational Awareness: A Systematic Review of the Literature. *Computers & Security* 46: 18–31.

Google Project Zero and Google DeepMind. 2024. *From Naptime to Big Sleep: Using Large Language Models to Catch Vulnerabilities in Real-World Code*. Available on-

line at https://projectzero.google/2024/10/from-naptime-to-big-sleep.html (Last accessed: September 6, 2026).

Help Net Security. 2025. *European Vulnerability Database Goes Live, But Who Benefits?* Available online at https://www.helpnetsecurity.com/2025/05/14/enisa-european-vulnerability-database-euvd/ (Last accessed: September 6, 2026).

Infosecurity Magazine. 2026. *NIST Drops NVD Enrichment for Pre-March 2026 Vulnerabilities*. Available online at https://www.infosecurity-magazine.com/news/nvd-enrichment-premarch-2026/ (Last accessed: September 6, 2026).

Jacobs, J., Romanosky, S., Edwards, B., Adjerid, I. and Roytman, M. 2021. Exploit Prediction Scoring System (EPSS). *Digital Threats: Research and Practice* 2 (3): 1–17.

Jajodia, S., Liu, P., Swarup, V. and Wang, C. (Eds) 2010. *Cyber Situational Awareness: Issues and Research*. New York, NY: Springer.

Mandiant. 2026. *M-Trends 2026: Data, Insights, and Strategies From the Frontlines*. Google Cloud. Available online at https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2026 (Last accessed: September 6, 2026).

Ministry of Defence (UK). 2016. *Joint Doctrine Publication 04: Understanding and Decision-making* (2nd Edn.). Shrivenham: Development, Concepts and Doctrine Centre.

NIST. 2026. *NIST Updates NVD Operations to Address Record CVE Growth*. National Institute of Standards and Technology. Available online at https://www.nist.gov/news-events/news/2026/04/nist-updates-nvd-operations-address-record-cve-growth (Last accessed: September 6, 2026).

Parasuraman, R. and Manzey, D.H. 2010. Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors* 52 (3): 381–410.

Recorded Future. 2026. *NIST NVD Enrichment Policy Change: Prioritizing Vulnerabilities with Attacker Behavior Signals*. Available online at https://www.recordedfuture.com/blog/nist-nvd-enrichment (Last accessed: September 6, 2026).

Verizon. 2026. *2026 Data Breach Investigations Report*. Basking Ridge, NJ: Verizon.

DISCLAIMER:
This text was drafted in a personal capacity. The views and opinions of the author expressed herein should therefore not be interpreted as stating or reflecting those of his employers or any organization of which he is a member.

THE INVISIBLE INFRASTRUCTURE OF CYBERSECURITY GOVERNANCE AND THE ROLE OF THE COOPERATIVE CYBER DEFENCE CENTRE OF EXCELLENCE (CCDCOE)

Ágnes Kasper
NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE), Estonia

Abstract
Cybersecurity governance is traditionally conceptualized through its highly tangible and visible components, such as technical architectures, legislative strategies, and specific operational capabilities. Drawing a conceptual parallel with air-traffic management, this chapter argues that these explicit structures are fundamentally enabled and sustained by a critical, yet frequently unexamined, "invisible infrastructure" of governance. This underlying framework consists of shared international normative understandings, common analytical methodologies, multi-disciplinary professional networks, and institutionalized mechanisms that translate high-level legal principles into daily operational practice. By examining these collaborative dynamics, the study underscores that international technological resilience and responsible state behavior cannot depend solely on isolated network boundaries, but require a structured alignment of strategic vocabulary, evidence collection standards, and shared professional trust.

1. Introduction

Cybersecurity governance is often described in solidly tangible and visible terms, such as technological capabilities, regulatory interventions and strategic objectives and preparedness – both in the civilian and military contexts. A prime practical example for such approach can be found in the European Union's 2020 Cybersecurity Strategy (European Commission and High Representative of the Union for Foreign Affairs and Security Policy, 2020), or even sharper in the NATO's cyber defence policy (NATO, 2021). However, arguably, cybersecurity governance is a more sophisticated system composed of interconnected layers, including also intangible elements and interactions among actors all striving to make an impact.

The most visible side of cybersecurity governance is akin to air-traffic management in an increasingly congested shared airspace. There is dense cross-border traffic between various categories of airfields and airports (IT systems, including hardware and software, technological capabilities), where effects of failures, miscalculations, or hostile acts can quickly propagate. International law provides the basic rules of navigation and responsibility, and while there is

no single global controller, the EU operates as a particularly integrated regional control area, where the NIS2, CRA, DORA, GDPR and other relevant legislation established detailed standards, reporting obligations, supervision, etc. The EU Commission, ENISA and related institutional and cooperation networks serve as control centers, exchanging information, assessing the performance of the system and designing corrective action if needed. States remain responsible for their national airspace, but cannot govern the risks alone and they remain dependent on their many counterparts and other elements in the system. Private sector builds and also owns much of the infrastructure, aircrafts, navigation systems, operates airlines, and floods passenger terminals (somewhat like social media platforms) with goods, services and manipulative ads, etc. Most research, policy interventions, normative reform, technological development work, training focuses on these tangible components.

However, these do not operate in isolation and behind them lies a much subtler layer of governance, like an invisible grid, where shared legal understanding, common analytical methods, professional networks, educational initiatives and mechanisms for translating legal principles into operational practice are forged, as suggested by Porter and Tan (Porter and Tan, 2023).

2. The Invisible Layer

When it comes to the international normative frameworks in the cyber context, the sometimes painstakingly slow and difficult progress in the UN demonstrates that the practical effect of the visible, formal rules depends on whether, or to what extent the people and institutions are connected and synchronized. For example, States may formally agree that international law, and in particular the UN Charter, "state sovereignty and international norms and principles that flow from sovereignty apply to State conduct of ICT-related activities, and to their jurisdiction over ICT infrastructure within their territory" (United Nations, 2013). However, they may still be unable to coordinate operationally if they do not have convergent vocabulary, approaches and analytical categories, understanding what conduct constitutes a violation of a rule, what forms of evidence may be used, when and what response options are triggered, or there are no trusted channels for exchanges. It is only a recent development that countries started to issue official statements on how – in their opinion – existing international rules apply to cyber activities, typically discussing their understandings on sovereignty, due diligence, non-intervention, use of force, international humanitarian law and state responsibility and attribution. At the time of writing there are 37 States and 2 international organizations (EU and the African Union) that have issued a national or common position (CyberLaw Toolkit, 2026).

Such complex and consequential statements, although there are significant differences among them, do not just pop-up overnight. The understanding, the motivation, the capacity, the content of such statements develops through long, non-public, informal and invisible processes, influenced by research, publications, consultations, conference discussions and similar interactions (Mačák, Dias and Kasper, 2025). Just like an electrical grid, the invisible grid of cybersecurity governance allows information, interpretation and assistance to move across its national and institutional components, aiming not at uniformity, but sufficient interoperability and synchronization to sustain collective action – enable the airports, control centers, aircraft manufacturers and others to perform their functions – or at least prevent major blackouts.

The invisible infrastructure of cybersecurity governance is supported by various enabling institutions, which form, maintain and strengthen this grid in the background. Such institutions can be universities, think-tanks, policy institutions, research services, centres of excellence, policy and government affairs in technology giants – form can vary, however what characterizes them is that they create knowledge, take the risk of thinking outside of the box, experiment, research, build capacity and create and maintain connections. This work is naturally riddled with failures, but both in its failures and successes enables institutions to interpret, operationalize and continuously adapt to current needs.

Cybersecurity governance is not controlled by a single actor. No institution, even such as the UN, NATO, the EU or CCDCOE, can simply design or engineer the governance system. Instead, the normative frameworks and governance structure emerge from the interactions of autonomous actors (Vivo, Katz and Ruhl, 2024), where institutions are governance nodes performing more subtle influencing, facilitating, coordinating, enabling and stabilizing functions. Accordingly, no single organization controls the development of international law in cyberspace but it emerges through repeated interaction of actors (Winston, 2023). This is facilitated by institutions, which create the conditions for international norms to adapt (Koontz et al., 2015). This chapter invites the reader to look beyond the visible elements of cybersecurity governance and illustrates the key functions of nodes in the invisible layer – i.e. in the 'knowledge infrastructure' – through the example of the CCDCOE.

3. Functioning of Knowledge Infrastructure in Cybersecurity Governance

Cybersecurity has been labelled as a protoscience (Hatleback, 2018). This view distinguishes the practice of cybersecurity from the science of cybersecurity, and accepts that substantial empirical knowledge has already been accumulated and there are even subfields, such as cryptography, with sophisticated

theories and rigorous methods. However, there are still no overarching unified principles, methods and laws for cybersecurity as a whole discipline, while it is moving towards this end. Just like alchemy became chemistry through systematic experimentation (Hatleback, 2018). Cybersecurity governance, as anchored also in governance and law, not only in technical cybersecurity, adds to the complexity.

Institutions, such as universities, think-tanks, research services contribute to consolidating knowledge and reducing complexity that comes along with systematic experimentation. These institutions in the knowledge infrastructure increase the **variety** of actors, which can influence how cybersecurity governance is shaped, including how international norms relating to cyber activities crystallize, emerge, apply or even loose relevance. For example, the Cybersecurity Master programme was established at Tallinn University of Technology in 2008, in accordance with the policy to do so indicated in Estonia's first Cybersecurity Strategy (Estonian Ministry of Defence, 2008), after the impactful 2007 cyberattacks against Estonia (Ottis, 2008). One of the main purposes of this move was to train specialists, however it is not an accident that this task was assigned to a university allowing safe experimentation. The new programme was a combination of technical and social sciences subjects, at that time a unique approach likely guided by the use-level experience of reacting to the large-scale cyberattacks in 2007 in technical, legal, political terms. The university started to yearly train a group of approx. 60 local and international students in relevant subjects, starting from system administration, foundations of cybersecurity, malware analysis, large-scale cyberattacks, through internet psychology to legal aspects of cybersecurity. By today the programme produced many leading experts contributing to the fabric of knowledge in the field with increased awareness about governance and legal challenges, both in the pubic and the private sector and academia, and it served as a model for other similar programmes elsewhere. Arguably, the designers, instructors and alumni of the programme shaped the way cybersecurity is understood not as a merely technical discipline, but one with just as much social relevance.

In the same year, the NATO CCDCOE was also formally established in Tallinn as an interdisciplinary knowledge hub, focusing on research, training and exercises throughout the fields of technology, strategy, operations and law (Loorents, Ottis and Rikk, 2009). This is not to say that there was a clear intention to give a push the protoscience of cybersecurity to become a full-fledged science in philosophical terms. Both steps were practical and use-oriented – for example NATO's policy for centres of excellence (COEs) emphasizes knowledge and expertise as the key asset of COEs and their value for the member nations (NATO, n.d.). It should also be noted that COEs do not have operational mandate (i.e. the NATO CCDCOE cannot and does not conduct cyberoperations),

hence COEs truly sit on the boundaries of policy and science of their subject matter area. COEs multiply the benefits for their members, the produced new research and insights become available for all instead of remaining confined by the limits of individual efforts.

Winston (2023) points out the importance of the **density** of relationships through which all the actors related to cybersecurity governance are connected. The more connections there are both within the knowledge infrastructure between actors, and between the knowledge infrastructure actors and actors, decision-makers in the visible layer of cybersecurity governance, the more opportunities there are for information, new knowledge and insights to move from one actor to another and influence the outcomes. COEs, and as such also the NATO CCDCOE, work with other international organizations, academia, governments and private industry. Experts working in the CCDCOE are seconded there by a contributing nations and they maintain their connections with their home networks, while serving also as a connection point with the networks of other experts, as well as the overall network of the CCDCOE. The resulting intense network of interrelations allows the CCDCOE to use multiple pathways to produce use-inspired research anchored in scientific theories and rigorous methods, achieve significant "buy-in" from actors in its experimentations, and reach to the visible, tangible layers of cybersecurity governance and operational practice via its outputs.

The current normative framework regarding cyber activities relies on the application of existing international law rules that date back decades or sometimes even centuries before the cyber era, and it is the voluntary and non-binding international norms that are cyber-specific. This points to need for learning and adaptation by communities, which requires **openness**. On the one hand, openness toward the technological changes and recognition that the use of information and communication technologies (ICTs) by States and non-state actors in the context of international security raises new challenges. On the other hand, openness toward learning and changes, building new capacities to respond to the new challenges.

The influence and implications of ICTs has been on the agenda of the United Nations since 1998, when the Russian Federation introduced a draft, leading to the adoption of resolution nr A/53/70 on 4 December 1998 by the General Assembly (United Nations General Assembly, 1999). It institutionalized an annual review of information security within the UN First Committee (Disarmament and International Security) and was followed by the establishment of a Group of Governmental Experts (GGE) in 2004 mandated to examine challenges in the field, which led to the formulation of norms of responsible state behavior in cyberspace in 2013 (Osula and Kasper, 2025), also putting forward the landmark position that "international law and in particular the [UN]

Charter, is applicable and is essential to maintaining peace and stability and promoting an open, secure, peaceful and accessible ICT environment" (United Nations, 2013). From 2018 the dialogue transitioned to a more inclusive global forum, the Open-Ended Working Group, paving the way for a single-track permanent UN cyber body – Global Mechanism on ICT Security –, which held its inaugural session in March 2026 (United Nations, 2026).

The above demonstrates the openness of States for dialogue, reflecting on how ICTs challenge established normative frameworks. It also shows both broadening and deepening understanding that normative frameworks need adaptations, however, the *acquis* accumulated in the UN forum is a less representative example of demonstrating openness to change by concrete action in the field of international law. The final OEWG report in 2025 drew strong criticism for its lack of ambition to push more the States to clarify how international law applies to cyber activities – which has been the core question since 2013, when the community agreed that international law does apply (Mačák, 2025).

4. Role of Institutions in the Knowledge Infrastructure – Legal Perspectives of Cybersecurity Governance

Against this backdrop, it should be noted that the core question – how international law applies in the cyber context – is far from simple, and early discussions of jus ad bellum and jus in bello concerning cyber attacks took place already around the turn of the millennium (Schmitt, 1999; Schmitt, 2002). However, after the 2007 cyber attacks against Estonia international law questions gained renewed attention, and the same overall question has been guiding the international law-oriented efforts of the CCDCOE from its establishment.

: In one of the earliest law outputs by the CCDCOE in 2010, Dion pointed out that interactions among international cyber incident responders were based on technical, operational, diplomatic and political relationships, not legal relationships (Dion, 2010). She reasoned that since most exiting international legal frameworks are not cyber-specific, they may be inapplicable or inefficient to properly address cyber incidents, however, she also observed that "before any new constructs or new laws are created, existing legal frameworks should be assessed to determine their appropriateness for managing global and international cyber threats" (Dion, 2010). In the same year, Häuβler examined the Articles 4 and 5 of the North Atlantic Treaty in different cyber attack scenarios and concluded that the legal consensus-building process needs innovative analysis challenging conventional wisdom (Häuβler, 2010). CCDCOE law researchers analyzed four international cyber incidents from 2007 and 2008, including the cyber attacks against Estonia, however their conclusions raised more questions

than answered (Tikk, Kaska and Vihul, 2010). In 2010 the Conference on Cyber Conflict (CyCon) was established, including a dedicated law and policy track, which today provides one of the most important forums to discuss in-depth legal issues related to collective cyber defence.

In this uncertain normative environment the groundbreaking project of the Tallinn Manual was launched in 2009, where the NATO CCDCOE invited an independent expert group to produce a manual on the law governing cyber warfare (Schmitt, 2013). It defined two threshold questions: whether the existing law applies at all and if yes, how?

On the first question, the Tallinn Manual was quick to refer to the Nuclear Weapons Opinion of the International Court of Justice, para. 39, which states that international law applies to "any use of force, regardless of the weapons employed", and the 2011 statement by the United States in its International Strategy for Cyberspace that cyberspace, despite its unique attributes, does not require reinvention of laws nor it renders existing ones obsolete.

The main aim of the Tallinn Manual project was to bring "some degree of clarity to the complex legal issues surrounding cyber operations" (Schmitt, 2013: 18). In an expert-driven process it engaged deeply with the second question, how international law applies, and provided guidance to actors concerned about the resulting ambiguities. As to the scope, the Tallinn Manual 1.0 generally focused on *jus ad bellum* and *jus in bello* matters. The follow-on project, Tallinn Manual 2.0, published in 2017, however, expanded on the highly influential first edition by extending its scope also to peacetime legal regimes (Schmitt, 2017). In both manuals the experts identify 'black letter rules' (95 and 154 respectively) and provide extensive commentary to each. About the authority of the Tallinn Manuals it is essential to understand that they are not official documents, but reflect the views of the contributors acting in their personal capacity. However, perhaps more important aspect than the authority of the Manuals, the drafting and reviewing process brought together experts from academia, technical, operational communities, as well as state representatives from nearly 50 nations (for Tallinn Manual 2.0). This can be understood as a form of co-production, even if the comments and contributions from state representatives cannot be identified in the text. Due to their process design, the Tallinn Manuals have reached the decision-maker layer even before the manuscript completion and publication.

Some of the greatest merits of the Tallinn Manuals is that they reduced the complexity of assessing whether international rules are appropriate for managing international cyber threats. It is not to say that they provide a final resolution on the issue, that was not the intention and such decisions are reserved for States. However, their analysis demonstrated that there are plausible and logical perspectives on how existing international law applies to cyber activities.

The Tallinn Manual 2.0 remains by far the most comprehensive study covering topics from general international law (including sovereignty, due diligence, jurisdiction, law of international responsibility), through specialized peacetime regimes (such as human rights, diplomatic and consular relations, law of the sea and air, space law), to international peace and security and international humanitarian law. This systematic and extensive analysis helps actors to make sense of the overwhelming amount of information, guiding them on what matters, what is relevant, what knowledge is reliable, which analytical methods should be used, etc. Moreover, it works toward shared legal understandings, reducing the complexity of legal reasoning in a technical area by providing a catalogue of possible interpretations of the legal rules. The Tallinn Manuals include views on which the group of international experts agreed, but also elaborated on issues where they were split and minority opinions appeared.

While in 2013 consensus emerged among States that international law applies in the cyber context, for many years to come States themselves, with occasional exceptions, still remained silent on *how* these rules apply. However, following the publication of Tallinn Manual 2.0 a clear trend started, and States begun to articulate their views publicly in official statements and positions and by today the clarification and development of the law in this area occurs to a great extent through these (Mačák, Dias and Kasper, 2025: 13). 2024 marked a turning point, and with the publication of the common positions of the African Union (African Union, 2024) and the EU common understanding (European Union and its Member States, 2024), now more than 100 states either have an individual position or joined a common one. While the motivations for developing a national or common position can be various, arguably the Tallinn Manual projects acted as a catalyst for these processes. Nevertheless, the impact of the Tallinn Manuals is unmistakably evidenced by adoption of their formulations and direct references to it by several national positions (e.g. positions of Germany, France, The Netherlands and Japan).

A comprehensive list of national and common positions is maintained in the CyberLaw Toolkit (https://cyberlaw.ccdcoe.org/wiki/Main_Page), another well-known and impactful project by an international consortium[1] with the participation of the NATO CCDCOE. (CyberLaw Toolkit, 2026) The CyberLaw Toolkit is an interactive environment, available fee of charge, providing access to relevant materials, hypothetical scenarios of cyber incidents and their legal assessment in the context of international law. As a capacity building resource in international cyber law (and excellent education & training material), the

[1] Six partner institutions: Czech National Cyber and Information Security Agency (NÚKIB); International Committee of the Red Cross (ICRC); NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE); University of Exeter, United Kingdom; U.S. Naval War College, United States; Wuhan University, China.

Toolkit is included in the database of the Global Mechanism on ICTs in the Context of International Security at the UN Office of Disarmament Affairs. : In 2025 the CyberLaw Toolkit has received the prestigious Jus Gentium research award from the American International Law Association (American Society of International Law, International Legal Research Interest Group, 2025). Managed as a common effort by the University of Exeter, NATO CCDCOE, and the National Cyber and Information Security Agency of the Czech Republic (NUKIB) since 2018, the Toolkit provides further systematization and real-life context for actors to make sense and sort through the overwhelming amount of information. Its nearly 40 peer-reviewed scenarios not only illustrate the application of existing law, but also serve as basis for discussions on the implications of the different interpretations.

While the Tallinn Manuals and the CyberLaw Toolkit primarily focus on substantive rules of international law, the Handbook on Developing a National Position on International Law and Cyber Activities (Mačák, Dias and Kasper, 2025) aimed to tackle the less salient issue related to processes shaping it. Having observed the growing number of national positions, the questions arose about how a State develops a national position, what are the motivations, triggers, aims, how it gets on the agenda, how it is initiated and organized, who leads and who is the penholder, who are consulted and when, what are the sources and methods used for analysis, how it is decided what topic is included and what is left out, why to use one format or another, how to present the resulting position, etc. The project core partners included the leading University of Exeter, CCDCOE, the Ministry of Foreign Affairs of Japan and the Ministry of Foreign Affairs of Estonia, and it was supported by institutions such as the Organization of American States, the Centre for International Law at the National University of Singapore, the African Union, the German Federal Foreign Office and the German Agency for International Cooperation, the Tallinn University of Technology, and it was funded by United Kingdom's Economic and Social Research Council Impact Acceleration Account (ESRC IAA). During the project three regional roundtables were organized in Washington DC, Singapore and Addis Ababa, bringing together 46 States and more than 70 experts for discussing challenges and exchanging good practices on developing a national position on how international law applies to cyber activities. The Handbook was launched in 2025 at CyCon in Tallinn, and it was also introduced in a side-event of the UN OEWG final session in New York. Now also translated into French and Spanish, the Handbook offers a practical and structured approach for all States developing or reviewing a national position, helping to foster greater legal clarity, predictability, and stability in cyberspace. By outlining existing practices, shared challenges, and strategic considerations, it offers a key resource to governments, legal practitioners, and policymakers

navigating the application of international law in the cyber context (Mačák, Dias and Kasper, 2025: 11).

The overall guiding question of *how* international law applies to cyber activities also includes the practical aspect of how they are operationalized. While the national and common positions most often use abstract formulations, the question remains how these rules are actually used in practice. This requires training, experimentation and learning. To facilitate these processes, the CCDCOE organizes Locked Shields, the worlds' largest cyber defence exercise annually, which includes parallel and integrated technical, legal and strategy tracks. In essence the Locked Shields is a complex real-time simulation spanning many days, where the defending teams from 40 countries face about 8000 cyber attacks on various system, including military systems and critical infrastructures, prompting also questions on the lawfulness of the conduct of different actors and their own response measures. This typically involves carefully designed challenges (injects) within the fictional overall scenario, where the legal advisors need to interact and cooperate with technical experts, and deliver clear and comprehensive legal advice to the military commander under time pressure and with incomplete information. Responses are scored and detailed feedback is provided to the participating teams. Injects are often prompted by real events reported by media outlets or encountered by the organizing team in their practice. Locked Shields allows two-way information flow between practitioners from 40 countries and the CCDCOE's team. First, through the careful design of injects and feedback to responses, existing legal approaches can be solidified and new situations or problems can be brought to the attention of the learning audiences. Second, the received responses not only indicate areas where more training is needed, but also areas where legal interpretations are converging, diverging or new perspectives are emerging.

Guided by the question of *how* international law applies, one needs to look forward, anticipate emerging challenges and prepare governance for changes. Much of this work is concentrated in the CyCon conference and the Annual Symposium on Cyber and International Law[2] and which are major platforms for exchanging new ideas, new agendas, new perspectives on international law in the cyber context. However, some of the most recent project of the CCDCOE highlight other pathways through which international norms and governance can be shaped. Since the CCDCOE is the custodian of the NATO cyber operations doctrine, it should be natural in the currently ongoing revision of the Allied Joint Doctrine for Cyberspace Operations (AJP 3.20) to take into account the

[2] Co-organized by American University Washington College of Law; Lieber Institute, West Point; The Federmann Cyber Security Research Center – Cyber Law Program; Hebrew University of Jerusalem, Centre of Excellence for National Security, NATO CCDCOE; the Centre for International Law, National University of Singapore, and the Oxford Process.

convergences identified in the positions of NATO allies on how international law applies to cyber activities. Since now the majority of NATO Member States either have an individual national position or joined a common understanding, it was possible to map out the convergence areas or a 'common baseline' within NATO, offering value in legal interoperability and doctrinal development (Kasper and Veersalu, 2026). Therefore this pathway serves the consolidation of international law and its interpretations in the national positions, and facilitates the uptake by the operational and policy communities – and breaks out from the recurrent cycle of international lawyers talking only to international lawyers.

Finally, the most recent and ongoing efforts also hold great promise related to frontier technologies. In the forthcoming edited volume "International Law and Artificial Intelligence in Armed Conflict: The AI-Cyber Interplay" (Mačák, Wobma and Tolppa, forthcoming), convened by the CCDCOE in cooperation with University of Exeter, the project combines the various functions and tools outlined above. By creating a common vocabulary and expert analysis at the point where law, technology and operations meet, the project goes beyond tracking of developments and aims to shape the dialogue. It's operational perspective makes it teachable in courses and exercise-ready for Locked Shields and other exercises, and with its practical focus on live questions it supports national and NATO policy work. The book's 19 substantive chapters revolve around the central theme whether international law still works in practice once AI and cyber reshape the speed, scale and architecture of military action. In its synthesis, the book identifies five structural stresses: temporal, epistemic, infrastructural, institutional and normative. It concludes that "the risk is not legal collapse but functional hollowing: the rules stay intact while their practical effect erodes" (Wobma and Tolppa, forthcoming). In short, it is a practical legal reference point for how AI is changing armed conflict – and above all for what happens where AI and cyber operations converge. Looking into the stressors on international law, the project anticipates challenges, supports adaptation and prepares the governance and legal institutions for changes.

5. Conclusions

This chapter invited the reader to look beyond the visible elements of cybersecurity governance and illustrated the key functions of nodes in the invisible 'knowledge infrastructure' using the example of the NATO CCDCOE. The illustrative examples highlighted the enabling role of institutions in the knowledge infrastructure of global cybersecurity governance and shaping international norms – distinguishing output from function. Not all research reaches the tangible layers of cybersecurity governance and its not always research that delivers an impact and shapes

policy decisions and normative frameworks. The picture is more complex and international norms, and their interpretations, emerge from the interaction of the various actors – including decision-makers and the less visible enabling institutions. Knowledge and information move along pathways between the actors (high-density relations provide more opportunity for knowledge and information to move) and there needs to be an openness in the community to be influenced, learn and adapt.

Enabling institutions might produce knowledge by researching how international law applies to cyber operations. However, when institutions engage in knowledge brokerage (Bandola-Gill, 2023), they catalyze the movement of knowledge and information between the actors when they bring government legal advisers, academics and operators together; synthesize national positions; develop common terminology; translate academic debates into operationally meaningful questions; and feed problems identified by practitioners back into research. CCDCOE is designed to both produce and broker knowledge, promoting not only the excellence of research but also its applicability.

CCDCOE projects illustrated some of the relevant functions of enabling institutions in the knowledge infrastructure. Tallinn Manuals project showed that *reducing complexity* and *co-production* are cornerstones for opening up the cyber governance community for changes and getting the ball rolling. The CyberLaw Toolkit *tracks* information (e.g. national positions) and offers elaborate and specialized scenarios that are used for *capacity building* across trainings and deliberations. The annual conferences, CyCon organized by the CCDCOE, and the Annual Symposium on Cyber & International Law co-organized by CCDCOE serve to *interface* and *connect* the actors in the community, including academia, operational community and decision-makers. CCDCOE also serves as an *experimentation* platform and *facilitates* learning by capturing experience and disseminating lessons learned in its annual Locked Shields cyber defence exercise (NATO Cooperative Cyber Defence Centre of Excellence, 2026) and similar events. With legal support to the NATO Cyber Operations Doctrine common understanding on how international law applies in the cyber context can be *consolidated* and *stabilized*. Finally, with its most recent project on International Law and Artificial Intelligence in Armed Conflict: The AI-Cyber Interplay, CCDCOE performs to *anticipate* emerging challenges and *preparing governance*, including normative frameworks, for adaptation and changes.

The case of the CCDCOE is a good example how an enabling institution in the invisible knowledge infrastructure of cybersecurity governance not only produces research and knowledge, but how it creates and maintains the knowledge ecosystem within which better research and better-informed decisions become possible. That said, many other institutions do the same or similar, and rather the question is how to cooperate, collaborate and how to achieve complementarities to build trust, networks, shared vocabularies, common understandings, open up new pathways for impact and build capacity to generate and use knowledge.

References

African Union. 2024. *Common African Position on the Application of International Law to the Use of Information and Communication Technologies in Cyberspace*. Available online at https://papsrepository.africanunion.org/server/api/core/bitstreams/65bdfced-80d9-445b-b57f-31037616ccda/content (Last accessed: November 7, 2026).

American Society of International Law, International Legal Research Interest Group. 2025. The Recipients of the Jus Gentium Award. Available online at https://asil.org/interest-group/international-legal-research/ (Last accessed: November 7, 2026).

Bandola-Gill, J. 2023. Knowledge brokering repertoires: Academic practices at science-policy interfaces as an epistemological bricolage. *Minerva* 61 (1): 71–92. https://doi.org/10.1007/s11024-022-09478-5.

CyberLaw Toolkit. 2026. National Positions on the Application of International Law to Cyber Activities. Available online at https://cyberlaw.ccdcoe.org/wiki/Main_Page (Last accessed: November 7, 2026).

Dion, M. 2010. *Different Legal Constructs for State Responsibility*. Tallinn: NATO Cooperative Cyber Defence Centre of Excellence.

Estonian Ministry of Defence. 2008. *Cyber Security Strategy 2008–2013*. Tallinn: Estonian Ministry of Defence.

European Commission and High Representative of the Union for Foreign Affairs and Security Policy. 2020. *The EU's Cybersecurity Strategy for the Digital Decade*. Joint Communication to the European Parliament and the Council, JOIN(2020) 18 final, 16 December.

European Union and its Member States. 2024. *Declaration on a Common Understanding of the Application of International Law to Cyberspace*. Council of the European Union, Doc. 15833/24, 18 November. Available online at https://www.consilium.europa.eu/en/press/press-releases/2024/11/18/cyberspace-council-approves-declaration-to-promote-common-understanding-of-application-of-international-law/ (Last accessed: November 7, 2026).

Hatleback, E.N. 2018. The protoscience of cybersecurity. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology* 15 (1): 5–12. https://doi.org/10.1177/1548512917737635.

Häußler, U. 2010. "Cyber Security and Defence from the Perspective of Articles 4 and 5 of the NATO Treaty". In C. Czosseck and K. Podins (Eds), *International Cyber Security: Legal & Policy Proceedings 2010*. Tallinn: Cooperative Cyber Defence Centre of Excellence, pp. 100–129.

Kasper, A. and Veersalu, K. 2026. *International Legal Perspectives on Cyber Operations: A Common Baseline for NATO Allies*. Tallinn: NATO Cooperative Cyber Defence Centre of Excellence. Available online at https://ccdcoe.org/library/publications/international-legal-perspectives-on-cyber-operations-a-common-baseline-for-nato-allies/ (Last accessed: November 7, 2026).

Koontz, T.M., Gupta, D., Mudliar, P. and Ranjan, P. 2015. Adaptive institutions in social-ecological systems governance: A synthesis framework. *Environmental Science & Policy* 53 (B): 139–151. https://doi.org/10.1016/j.envsci.2015.01.003.

Loorents, P., Ottis, R. and Rikk, R. 2009. "Cyber Society and Cooperative Cyber Defence". In N. Aykin (Ed.), *Internationalization, Design* (LNCS 5623). Berlin: Springer, pp. 180–186.

Mačák, K. 2025. A missed opportunity: An international law perspective on the final OEWG report. *EJIL: Talk!*, 22 July. Available online at https://www.ejiltalk.org/a-missed-opportunity-an-international-law-perspective-on-the-final-oewg-report/ (Last accessed: November 7, 2026).

Mačák, K., Dias, T. and Kasper, Á. 2025. *Handbook on Developing a National Position on International Law and Cyber Activities: A Practical Guide for States*. Tallinn: NATO Cooperative Cyber Defence Centre of Excellence. Available online at https://ccdcoe.org/library/publications/handbook-on-developing-a-national-position-on-international-law-and-cyber-activities-a-practical-guide-for-states/ (Last accessed: November 7, 2026).

NATO. 2021. *Comprehensive Cyber Defence Policy*. Endorsed by NATO Heads of State and Government at the Brussels Summit, 14 June (not publicly released); see NATO, *Brussels Summit Communiqué*, 14 June 2021, para. 32.

NATO. n.d. *Final Decision on MC 0685: Military Committee Policy for Centres of Excellence*. Not publicly released.

NATO Cooperative Cyber Defence Centre of Excellence. 2026. Exercises. Available online at https://ccdcoe.org/exercises/ (Last accessed: November 7, 2026).

Osula, A.-M. and Kasper, A. 2025. "International Law and Cyberspace". In G. Christou, W. Vosse, J. Burton and J.A. Koops (Eds), *The Palgrave Handbook on Cyber Diplomacy*. Cham: Palgrave Macmillan, pp. 77–101. https://doi.org/10.1007/978-3-031-93385-1_4.

Ottis, R. 2008. "Analysis of the 2007 Cyber Attacks Against Estonia from the Information Warfare Perspective". In *Proceedings of the 7th European Conference on Information Warfare and Security*. Reading: Academic Conferences Limited, pp. 163–168.

Porter, T. and Tan, N. 2023. An integrated complex adaptive governmental policy response to cyberthreats. *Journal of Economic Policy Reform* 26 (3): 283–297. https://doi.org/10.1080/17487870.2022.2125390.

Schmitt, M.N. 1999. Computer network attacks and the use of force in international law: Thoughts on a normative framework. *Columbia Journal of Transnational Law* 37: 885–937.

Schmitt, M.N. 2002. Wired warfare: Computer network attacks and jus in bello. *International Review of the Red Cross* 84: 365–399.

Schmitt, M.N. (Ed.) 2013. *Tallinn Manual on the International Law Applicable to Cyber Warfare*. Cambridge: Cambridge University Press.

Schmitt, M.N. (Ed.) 2017. *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. Cambridge: Cambridge University Press.

Tikk, E., Kaska, K. and Vihul, L. 2010. *International Cyber Incidents: Legal Considerations*. Tallinn: NATO Cooperative Cyber Defence Centre of Excellence.

United Nations. 2013. *Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security*. UN Doc. A/68/98, 24 June.

United Nations. 2026. *Provisional Agenda and Annotations: Global Mechanism on Developments in the Field of Information and Communications Technologies in the Context of International Security and Advancing Responsible State Behaviour in the Use of Information and Communications Technologies*, Organizational Session, New York, 30 and 31 March 2026. UN Doc. A/AC.304/2026/1, 19 March. Available online at https://digitallibrary.un.org/record/4106867 (Last accessed: November 7, 2026).

United Nations General Assembly. 1998. *Developments in the Field of Information and Telecommunications in the Context of International Security*. Resolution A/RES/53/70, 4 December.

Vivo, P., Katz, D.M. and Ruhl, J.B. 2024. A complexity science approach to law and governance. *Philosophical Transactions of the Royal Society A* 382: 20230166. https://doi.org/10.1098/rsta.2023.0166.

Winston, C. 2023. International norms as emergent properties of complex adaptive systems. *International Studies Quarterly* 67 (3): sqad063. https://doi.org/10.1093/isq/sqad063.

Wobma, N. and Tolppa, M. Forthcoming. "Concluding Synthesis: Structural Stresses and Emerging Recommendations". In *International Law and Artificial Intelligence in Armed Conflict: The AI–Cyber Interplay*. Oxford: Oxford University Press.

CYBERSPACE AND THE UNITED NATIONS: TOWARDS A GLOBAL FRAMEWORK FOR RESPONSIBLE STATE BEHAVIOUR

Giorgio Marrapodi
Permanent Representative of Italy to the United Nations in New York

Abstract
This Chapter is dedicated to understanding why the United Nations is addressing the topic of Cybersecurity. While it is obvious that the drafters of the UN Charter could not imagine, 80 years ago, how both international society and international politics would have evolved, it is nonetheless very clear that the aims of the UN – peace and security, development, promotion of human rights – were conceived from the outset as pillars which had to resist vis-à-vis any new challenge humanity had to face. Through a brief historical overview, on one side this Chapter examines the different steps that have led to the institution of the "Global Mechanism on developments in the field of ICTs in the context of international security and advancing responsible State behaviour in the use of ICTs". On the other side, it also focuses on the perspectives of this new exercise against the backdrop of a complex multilateral environment. The Chapter concludes with a focus dedicated to the role that Italy, as a prominent actor on the international arena and as a staunch promoter of multilateralism, plays in contributing to shape the debate on cyber issues at the UN, thus actively taking part to the way the International Community intends to address a question which is no longer to be considered the trademark of the future, but it is already one of the most defining aspects of our everyday life.

1. Why is the United Nations Addressing Cyber Security?

In his fundamental handbook on the United Nations, an Italian Professor of International Law, Benedetto Conforti, explains that the wide array of issues that the UN engages with, as enshrined in Article 1 of the Charter, makes this international organization "*political*" (Conforti, 2005) in its very essence.

Can we imagine anything more political than "cyberspace" (National Institute of Standards and Technology, n.d.)[1], or more generally, "cyber issues" in today's world? We are living in an era in which every single operation of our life is characterized by a high degree of digital content. Let alone the industrial

[1] An effective definition of cyberspace is the one provided by the United States National Institute of Standards and Technology, according to which cyberspace is "the interdependent network of information technology infrastructures, and includes the Internet, telecommunications networks, computer systems and embedded processors and controllers in critical industries". See https://csrc.nist.rip/glossary/term/cyberspace.

sector, which is *per se* founded on technological developments, the other two columns of the economy – agriculture and services sector – are also now inextricably dependent on the digital sphere. On one side, this evidence is a source of hope, as it envisages a process of "360 degree" development whose benefits concern the well-being of people. Let's just imagine how technology can make agriculture more resilient to natural disasters or more productive in terms of outcome; the same approach concerns the field of services, including those provided by public administrations, including diplomacy: a field in which digitalization has boosted the accessibility of services to the broader public, but also the transparency of processes and respect for the rights of customers. On the other side, the magnitude of risks and challenges is also considerable. As cyber space is not bounded by physical limits and is virtually accessible to everybody, everywhere and at any time, the number of possible shortcomings in its correct functioning for the benefit of human beings is significant. These shortcomings include malfunctioning, vulnerability due to external factors (natural events, blackouts, etc.) but also to malign actions by actors motivated by hostile aims, on a scale that is not comparable to that of the "pre-digital" world[2].

What started as a revolution is now a reality so deeply rooted that it has shaped the very nature of our society, with a deep impact on the way we live, interact with each other, and express ourselves as human beings.

If this challenge calls every State to act in order to guarantee that its citizens can enjoy the benefit of progress, while at the same time staying protected from its risks, the international community as a whole also needs to be involved, for the very fact that – as already explained – there are no geographical limits to cyberspace.

It is therefore clear that, if the United Nations is called, *inter alia*, "to achieve international co-operation in solving international problems of an economic, social, cultural, or humanitarian character", it cannot be exempted from dealing with the wide arrays of issues connected to the "cyber domain." And it cannot be exempted from doing so because it is a Universal Organization, whose membership covers every corner of the world, where the whole international community meets every day to achieve, through multilateral diplomacy, peace and security, respect for human rights, dignity and equality for all.

[2] Let's take a simple example: in the pre-digital era, a robbery necessarily involved an activity naturally bounded by several limits (e.g. the thief had to be physically present in the place he wanted to rob) and possibly had limited consequences (managing to break into a bank allowed the thief to rob only the assets kept in that specific place). Nowadays, a robbery can happen any time without the thief moving from his computer desk; breaking through the firewall of an online banking system can cause damage that goes far beyond a single, specific bank branch.

2. UN and Cyberspace. The Road to a Global Mechanism

Having said why the United Nations responds to issues related the cyber domain, we can now start to focus on how it does so.

It was at the beginning of the 1990s (Maurer, n.d.) that the United Nations started engaging in subjects connected with cyberspace, when the topic of the criminal use of emerging technologies started arising among the issues debated in the pertinent fora, such as UNODC (United Nations Office on Drugs and Crime). Those were the years in which the Internet was becoming little by little familiar to many households around the world, but its real potential was already clear in the scientific community and at the political level. It was therefore in 1998 that the Russian Government decided to introduce a draft resolution to be approved by the 1st Committee of the United Nations General Assembly, the one devoted to Disarmament and International Security, whose title was significantly, "Developments in the field of information and telecommunications in the context of international security" (United Nations General Assembly, 1998). This text paved the way for a series of Resolutions on the same topic, the last of which was adopted in 2023 (United Nations General Assembly, 2023)[3], in a totally different geopolitical and social environment.

The 1st Committee was the first forum to expressly address this subject, but it was not the only one. The 3rd Committee – Social, Humanitarian and Cultural Committee – followed suit with discussions on the issue from the point of view of crime prevention and criminal justice (United Nations General Assembly, 2000), as did 2nd Committee – Economic and Financial Committee – from the point of view of creating "a global culture of cyber-security." (United Nations General Assembly, 2002). Therefore, it was also time for the Economic and Social Council (ECOSOC) to address some of these issues, from the perspective of the impact of cyberattacks on development[4], as well as for other United Nations Agencies and Offices (Maurer, n.d.).

The momentum generated by Resolution 53/70 and by the increasing debate within the United Nations on cyber-related issues led to the subsequent establishment of several intergovernmental processes devoted to addressing the security of and use of ICTs in the context of international security (United Nations Office for Disarmament Affairs, n.d.). The first form in which this process developed was that of the Group of Governmental Experts, the first of which was established in 2004, followed by five others. All of them were dedicated to

[3] A/RES/78/237, adopted on 22nd December 2023 with a record of 104 in favour, to 53 against, with 7 abstentions.

[4] This was the case of the "special event on cybersecurity and development", held in New York on 9th December 2011.

studying the threats posed by the use of information and telecommunication technologies in the context of international security, and, of course, how to address them. The outcome of these Groups was generally the production of agreed reports containing recommendations for the United Nations membership (Dominioni & Persi Paoli, 2026).

The GGEs adopted reports in 2010, 2013, 2015 and 2021. The last of these was the outcome of the sixth and final GGE, which met between 2019 and 2021. The time was ripe to "upgrade" the way the United Nations addressed this topic. Therefore, in 2020 (United Nations General Assembly, 2020)[5], the General Assembly, through resolution 75/240, decided to establish a new five-year (2021-2025) Open-Ended Working Group on security of and in the use of information and communications technologies.

Very briefly, the main topics that both the GGEs and the OEWGs dealt with were: Existing and emerging threats; International law and the use of ICTs; Norms, rules and principles of responsible behaviour of States; Confidence-building measures; Capacity-building (United Nations General Assembly, 2022)[6].

As we will see in the next paragraph, these topics constitute the "pillars"[7] (Open-Ended Working Group, 2024) of the current structure of the "Global Mechanism on developments in the field of ICTs in the context of international security and advancing responsible State behaviour in the use of ICTs"[8].

[5] Indeed, a first "whole of UN" approach to the issues was adopted in 2018 with General Assembly Resolution 73/27, which established an Open-Ended Working Group (OEWG), open to all Member States, to begin its work in 2019 and hold intersessional consultative meetings with industry, civil society and academia. The Group adopted a report by consensus at its final session in March 2021 (A/75/816). The final report and the recommendations contained therein were endorsed in General Assembly decision 75/564.

[6] In 2022, with Resolution A/RES/77/37, entitled "Programme of action to advance responsible State behaviour in the use of information and communications technologies in the context of international security", the UN General Assembly requested for the first time a report of the Secretary-General, taking into account the views submitted by States, on the proposal, "to establish a United Nations programme of action to advance responsible State behaviour in the use of information and communications technologies in the context of international security, as a permanent, inclusive, action – oriented mechanism to discuss existing and potential threats; to support States' capacities and efforts to implement and advance commitments to be guided by the framework for responsible State behaviour, which includes voluntary, non-binding norms for the application of international law to the use of information and communications technologies by States, confidence-building and capacity building measures […]; to discuss, and further develop if appropriate, this framework; to promote engagement and cooperation with relevant stakeholders; and to periodically review the progress made in the implementation of the programme of action as well as the programme's future work" (A/RES/77/37, par. 1). This Report was subsequently issued on 18 April 2023, as UN doc. A/78/76.

[7] The five "pillars" of responsible State behaviour in the use of ICTs are detailed in Annex C, par. 9 of the 2024 Report of the open-ended working group on security of and in the use of information and communications technologies 2021-2025 (published as UN document A/79/214).

[8] The Open-Ended Working Group on security of and in the use of information and

3. The Global Mechanism. Opportunities and Challenges Ahead of this New Instrument

The Global Mechanism, the arrival point of the OEWG 2021-2025, started its work in 2026. It "aims to facilitate the sharing of information and best practices on cyber threats, voluntary norms of responsible behaviour by States in cyberspace, the applicability of international law and confidence-building measures, whilst supporting the collective response to emerging cyber threats, including through the development of capacity-building activities." (Italian Ministry of Foreign Affairs and International Cooperation, n.d.)

After an organizational meeting in March – aimed at defining some preliminary and procedural aspects – its first plenary session was convened during the summer, at the end of July 2026. The difference with all the previous formats that dealt with cyber issues in the United Nations, is that the Global Mechanism is a permanent forum, with a clear mandate which defines its structure, purposes and its very organization[9].

This is not a secondary aspect to consider: in a historical context in which it is getting more and more difficult for the United Nations to reach consensus on topics related to international security, having agreed on the functioning of this new mechanism is a rare and precious opportunity. But this does not mean that all the challenges have been tackled once and for all and that from now on it will be possible only to achieve further progress without steps backwards. On the contrary, the story of the Global Mechanism is far from fully written. On one side, looking at the medium-long term, it is impossible to predict whether the several tensions that characterize today's international arena will dissolve or harden. At the end of a five-year cycle, the Global Mechanism will have to face its first review conference. Will this appointment be successful? Will the Global Mechanism be adjusted according to the experience of its first cycle, or will it be made irrelevant? It is very difficult, if not impossible, to say today.

On the other side, looking at the very short term, the most important feature of the Global Mechanism, its "Dedicated Thematic Groups,"[10] will show

communications technologies, published its Final Report in July 2025. This document (https://docs.un.org/en/A/80/257) which was finally approved in December 2025 by United Nations General Assembly Resolution A/RES/80/16, led to the establishment of the 'Global Mechanism on developments in the field of ICTs in the context of international security and advancing responsible State behaviour in the use of ICTs'.

[9] See UN document A/80/257, Annex I "Additional Elements for the Global Mechanism on developments in the field of ICTs in the context of international security and advancing responsible State behaviour in the use of ICTs".

[10] "The work of dedicated thematic groups will aim to build on and complement the discussions in the substantive plenary sessions by providing the opportunity for more detailed and action-oriented discussions, drawing on the five pillars of the framework, in line with the

whether the delicate balance between a government-led process and the substantive involvement of a wide array of stakeholders will be able to work. In fact, it has been observed that "cyber diplomacy goes beyond the conventional margins of diplomacy. It involves non-state actors that include corporations and 'civil society'" (Lewis, 2026) – which is normal for a phenomenon strictly connected with every aspect of our life – but also that "no state will allow its interests to be safeguarded by technicians, executives, lawyers or academics" (Lewis, 2026) – which is also true and inevitable for an international community whose main actors remain the States. Even though it would be in everybody's interest to have a highly democratic process, in which the voices of all the relevant stakeholders are heard to shape future decisions and national positions on the basis of a deeper and more comprehensive understanding of all the issues at stake, we have to be realistic about the fact that this will hardly happen, at least not in the next few years. Not all international actors, in fact, are ready to be publicly confronted with views which can be different from theirs; to face criticism from actors that are not always perceived as unbiased or genuinely independent; or to risk a process which was intended to be mostly technical turning into a platform for political struggle.

This does not mean that the DTGs and the Global Mechanism will not work; more simply, it means that they are based on a delicate equilibrium which has to be nurtured through daily and patient diplomatic work.

Italy, also through the contribution of the Permanent Mission to the United Nations in New York, is ready to participate in this diplomatic effort. The five pillars of the Global Mechanism are each very relevant, and on each of them Italy has its own clear negotiating position,[11] built through the contribution of all the pertinent public administrations and continuous dialogue with different stakeholders. This position is carried forward in close coordination with EU partners and in dialogue with the *likeminded* States, in order to broaden the ability to engage in dialogue with the wider membership, with a true multilateral attitude.

For this same reason, a few additional words can be spent on the issue of capacity-building, one of the five pillars of the Global Mechanism, and one of particular interest to Italy[12]. In fact, it is thanks to it and through it that the

agreed functions and scope of the Global Mechanism" (see UN document A/80/257, Annex I, par. 6).

[11] Which will not be examined in this brief chapter, but which have been already detailed in official documents, such as the National Document on International Law and Cyberspace (https://www.esteri.it/wp-content/uploads/2021/11/Italian-Position-Paper-on-International-Law-and-Cyberspace.pdf) or the National Cybersecurity Strategy (https://www.acn.gov.it/portale/strategia-nazionale-di-cybersicurezza).

[12] Italy conducts several capacity-building programs, both bilaterally and through

entire international community will be able to actively take part in the work of the Global Mechanism and contribute to finding solutions to problems which concern the whole planet. Cybersecurity, in fact, is a domain which does not contemplate borders and does not distinguish between developed and developing Countries.[13]

4. Italy and the United Nations. Being a Protagonist also in the Field of Cyberspace

Italy has been a member of the United Nations since December 14, 1955. That day was just the beginning of a long history of collaboration, support and engagement within and for the United Nations. It was the most visible and logical consequence of the multilateral approach that Italian foreign policy has put at the core of national interests and that shapes the action of Italian diplomacy.

The same commitment with which Italy participates in the UN's activities to pursue the Charter's objectives, is at the foundation of Italy's approach to the Global Mechanism. With the same spirit that has led Italy over the decades to contribute with determination to the drafting of several General Assembly and Security Council resolutions that have made history[14], it is now time to work to ensure that the Global Mechanism bears good fruit for the benefit of the whole international community.

The method of work is the same as always: narrowing differences in positions and broadening consensus on many issues.

The approach will be to keep the focus on an international cooperation conducive to sharing information on cybersecurity, preventing cyberattacks, helping to build cyber capabilities in developing countries, and establishing a framework for the governance of cyberspace. In doing so, Italy's contribution

multilateral organizations, such as the World Bank, UNIDIR (United Nations Institute for Disarmament Research) and the International Telecommunications Union. Examples of projects supported by Italy range from developing national cybersecurity strategies to training personnel; from delivering information and communications technologies products to crafting awareness-raising campaigns.

[13] Cyber capacity-building is extremely relevant also in practical terms: it is an essential tool for strengthening national resilience against increasingly sophisticated cyber threats, to help safeguard critical infrastructure, protect national security and reduce dependence on external actors, to ensure that technological innovations are adopted securely. In terms of international cooperation, it also promotes trust and stability in cyberspace, while at the same time fostering sustainable digital development.

[14] Just to mention a few: the campaigns for a moratorium on the death penalty; for the rights of women and minors; against the practice of female genital mutilation and of early and forced marriages; against all forms of religious discrimination and in favour of freedom of expression.

to the work of the Global Mechanism is further enhanced by the participation of national public- and private-sector organizations operating in the field of cybersecurity (Italian Ministry of Foreign Affairs and International Cooperation, n.d.).

Probably more than any other multilateral forum, due to its universal membership and the simultaneous presence of delegates from every corner of the world, the United Nations is the best place for Italy to promote its views on cyber issues: a free, secure and accessible cyberspace that respects human rights and in which international law is fully applicable. A new challenge for Italian diplomacy, but also a unique opportunity to be a protagonist once again.

As a prominent actor in the international arena and as a staunch promoter of multilateralism, Italy has the responsibility, and the ability, to do so. By participating in shaping the debate on information and communication technologies, Italy will actively contribute to the way the international community intends to address a matter which is no longer to be considered the trademark of the future, but is already one of the most defining aspects of our everyday life.

References

Agenzia per la Cybersicurezza Nazionale. (2022). National Cybersecurity Strategy. https://www.acn.gov.it/portale/strategia-nazionale-di-cybersicurezza

Conforti, B. (2005). Le Nazioni Unite. CEDAM.

Dominioni, S., & Persi Paoli, G. (2026). Unpacking the United Nations Open-Ended Working Group on ICT in the Context of International Security (2021–2025). UNIDIR.

Italian Ministry of Foreign Affairs and International Cooperation. (n.d.). Cyber diplomacy. https://www.esteri.it/en/temi/diplomazia_cyber/diplomazia-cibernetica/

Italian Ministry of Foreign Affairs and International Cooperation. (2021). Italian position paper on international law and cyberspace. https://www.esteri.it/wp-content/uploads/2021/11/Italian-Position-Paper-on-International-Law-and-Cyberspace.pdf

Lewis, J. A. (2026). The practice of cyber diplomacy. In A. Salvi, H. Tiirmaa-Klaar, & J. A. Lewis (Eds.), A Handbook for Practice of Cyber Diplomacy (pp. 22–23). European Union Institute for Security Studies.

Maurer, T. (n.d.). Cybersecurity and the United Nations. Freedom Online Coalition. https://freedomonlinecoalition.com/blog/cybersecurity-and-the-united-nations-by-tim-maurer/

National Institute of Standards and Technology. (n.d.). Cyberspace. https://csrc.nist.rip/glossary/term/cyberspace

Open-Ended Working Group on security of and in the use of information and communications technologies. (2024). Final report of the Open-Ended Working Group

on security of and in the use of information and communications technologies 2021–2025 (A/79/214).

Open-Ended Working Group on security of and in the use of information and communications technologies. (2025). Final report of the Open-Ended Working Group on security of and in the use of information and communications technologies 2021–2025 (A/80/257). https://docs.un.org/en/A/80/257

United Nations General Assembly. (1998). Developments in the field of information and telecommunications in the context of international security (A/RES/53/70). https://docs.un.org/en/A/RES/53/70

United Nations General Assembly. (2000). Combating the criminal misuse of information technologies (A/RES/55/63). https://docs.un.org/en/A/RES/55/63

United Nations General Assembly. (2002). Creation of a global culture of cybersecurity (A/RES/57/239). https://docs.un.org/en/A/RES/57/239

United Nations General Assembly. (2018). Developments in the field of information and telecommunications in the context of international security (A/RES/73/27). https://docs.un.org/en/A/RES/73/27

United Nations General Assembly. (2021). Decision 75/564: Open-ended working group on developments in the field of information and telecommunications in the context of international security. https://docs.un.org/en/A/DEC/75/564

United Nations General Assembly. (2021). Report of the Open-ended Working Group on developments in the field of information and telecommunications in the context of international security (A/75/816). https://docs.un.org/en/A/75/816

United Nations Secretary-General. (2023). Developments in the field of information and telecommunications in the context of international security (A/78/76). https://docs.un.org/en/A/78/76

United Nations General Assembly. (2020). Developments in the field of information and telecommunications in the context of international security (A/RES/75/240). https://docs.un.org/en/A/RES/75/240

United Nations General Assembly. (2022). Programme of action to advance responsible State behaviour in the use of information and communications technologies in the context of international security (A/RES/77/37). https://docs.un.org/en/A/RES/77/37

United Nations General Assembly. (2023). Developments in the field of information and telecommunications in the context of international security (A/RES/78/237). https://docs.un.org/en/A/RES/78/237

United Nations General Assembly. (2025). Global Mechanism on developments in the field of ICTs in the context of international security and advancing responsible State behaviour in the use of ICTs (A/RES/80/16). https://docs.un.org/en/A/RES/80/16

United Nations Office for Disarmament Affairs. (n.d.). Developments in the field of information and telecommunications in the context of international security. https://disarmament.unoda.org/en/our-work/emerging-challenges/developments-field-information-and-telecommunications-context

CYBERDIPLOMACY: NORMS OF RESPONSIBLE STATE BEHAVIOR AND THE INTERNATIONAL COOPERATION FRAMEWORK

Diego Brasioli
Italian Ministry of Foreign Affairs and International Cooperation

Abstract
In an era of rapid digital transformation, cyberspace has become central to geopolitical interaction, economic development, and societal connectivity, while simultaneously increasing global vulnerability to malicious cyber activities. This article examines the evolution of cyberdiplomacy as a framework for promoting responsible state behavior and international cooperation in the digital domain. It analyzes the development of United Nations processes, including the articulation of voluntary norms affirming the applicability of international law, the importance of due diligence, and the protection of critical infrastructure. The study also highlights the role of confidence-building measures, capacity-building initiatives, and regional organizations such as the EU, OSCE, and ASEAN in operationalizing these norms. Addressing tensions between sovereignty and global interdependence, the article underscores the need for multi-stakeholder engagement, public-private partnership, and sustained political commitment. It concludes that cyberdiplomacy has evolved into a multi-level governance architecture aimed at strengthening stability, predictability, and resilience in cyberspace.

1. Introduction: Cyber-diplomacy and the Institutionalization of Cyber Governance

The rapid digital transformation of the past three decades has fundamentally reshaped economic development, public administration, and international security. Digital infrastructure now underpins energy systems, financial markets, healthcare networks, supply chains, and electoral processes. Joseph Nye has defined "cyber power" as «*the ability to use cyberspace to create advantages and influence events in other operational environments*» (Nye, 2011).

Digital security is not merely about protecting computer data, but rather about influencing military, economic, and regulatory balances through digital infrastructures. In the networked era, power is increasingly a function of connectivity. Cyberspace has therefore evolved from a technical communications medium into a strategic domain of geopolitical interaction. States increasingly recognize that cyber capabilities influence deterrence, economic competitiveness, and national resilience.

Yet growing interconnectivity has also generated systemic vulnerability.

Malicious cyber operations, ranging from state-sponsored intrusions to ransomware campaigns conducted by organized criminal groups, have disrupted critical infrastructure and democratic institutions while imposing severe economic costs. IBM's Cost of a Data Breach Report 2025 (IBM Security, 2025) estimates the global average cost of a breach at USD 4.44 million, rising to USD 10.22 million in the United States. Even more conservative academic estimates place annual global cybercrime losses between USD 500 billion and USD 1 trillion, while forward-looking assessments warn that the global economic impact of cyberattacks could reach USD 15 trillion annually by 2030 if current trends persist (Cyber Defense Wire, 2025).

The governance challenges of cyberspace are not without historical precedent. In *Mare Liberum* (1609), Hugo Grotius rejected claims of exclusive maritime sovereignty, arguing that the seas (particularly in their high-seas dimension) could not be appropriated and instead required shared legal principles to preserve openness and freedom of navigation (Grotius, 1609/1916). The high seas thus emerged as a non-appropriable space regulated through common legal norms rather than exclusive dominion.

Cyberspace presents a structurally related but legally distinct configuration. Unlike the oceans, which exist as a natural environment beyond ownership, cyberspace is constituted by physically located and legally regulated infrastructures. Although global in function, digital networks depend on servers, data centers, submarine cables, and exchange points situated within national territories and subject to domestic jurisdiction. The distinction is therefore not merely one of private operation, but of structural design: cyberspace is a constructed and territorially embedded network rather than a naturally occurring commons. It cannot be equated with a classical *res communis*, nor reduced to a fully sovereign enclosure. Instead, it represents an interoperable yet jurisdictionally fragmented domain in which sovereignty and interdependence coexist in persistent tension.

From an international law perspective, Grotius' relevance lies not in analogical transplantation but in normative structure. Shared domains require regulation without monopolization. UN processes reaffirm that existing international law, including the UN Charter, applies to state conduct in cyberspace, which in turn is thus integrated into the state-centered legal order rather than treated as an exceptional domain. In other words, what emerges is not a digital *mare liberum* in absolute terms, but a regulated domain: functionally open, yet stabilized through state responsibility, due diligence, and cooperative governance. As in Grotius' time, stability depends on balancing sovereignty with shared legal restraint in a space no state can fully control yet all can disrupt.

Compounding this complexity is the role of non-state actors. Ransomware groups, cyber mercenaries, and criminal networks operate across jurisdictions,

sometimes with tacit state tolerance. Their growing professionalization and convergence with geopolitical agendas further destabilize the digital environment (World Economic Forum, 2026).

In response, states have increasingly turned to diplomacy. The UN Group of Governmental Experts affirmed that international law applies to cyberspace and articulated voluntary norms of responsible behavior (UN GGE Report 2015). The Open-Ended Working Group broadened participation and emphasized confidence-building and capacity development (UN OEWG Report 2021). The 2026 establishment of a Permanent Mechanism marks a further step toward structural consolidation.

Cyber capacity building and public–private partnerships have become central to this architecture, underscoring the importance of national capability development and collaboration with the private sector that manages much of the global digital infrastructure. Cyberdiplomacy thus represents an emerging multi-level governance framework aimed at reducing uncertainty, mitigating escalation risks, and promoting responsible state behavior in an increasingly contested and economically consequential domain.

This article argues that cyberdiplomacy has evolved beyond norm articulation into a structured, multi-level implementation architecture embedded within a broader regime-complex structure characterized by polycentric implementation dynamics. The transition from the GGE to the OEWG and ultimately to a UN Permanent Mechanism illustrates the gradual institutionalization of cyber governance at the global level. Regional initiatives in the European Union, the OSCE, and ASEAN, alongside political coordination within the G7 framework, complement this process by operationalizing norms, strengthening resilience, and promoting coordinated responses.

2. Cyberdiplomacy in the United Nations: from the GGE and OEWG to the Permanent Mechanism: Institutionalization of ICT Security and the Introduction of Two Dedicated Thematic Groups

United Nations engagement on cybersecurity has evolved through distinct but interconnected institutional phases over the past two decades. The process formally began in 2004, when the General Assembly adopted resolution A/RES/58/32, establishing the first Group of Governmental Experts (GGE) on developments in the field of information and telecommunications in the context of international security (United Nations, 2004). Conceived as a limited, expert-driven forum, the GGE brought together a small number of appointed states tasked with building consensus on principles governing state conduct in cyberspace. This format prioritized technical deliberation and diplomatic flex-

ibility but operated within structural constraints of limited participation and time-bound mandates.

The GGE

The 2013 and 2015 GGE reports marked a watershed in international cyber governance. They affirmed that existing international law applies to cyberspace and articulated eleven voluntary, non-binding norms of responsible state behaviour (United Nations, 2013; United Nations, 2015). These norms addressed the protection of critical infrastructure, cooperation through national Computer Emergency Response Teams (CERTs), and the responsibility of states to prevent malicious ICT activity emanating from their territory. Despite these achievements, the breakdown of consensus during the 2016–2017 GGE cycle exposed the fragility of a process dependent on limited participation.

The OEWG

In response, the General Assembly adopted resolution A/RES/73/27, establishing the Open-Ended Working Group (OEWG) and expanding participation to all Member States (United Nations, 2018). This shift represented a structural recalibration toward inclusivity and transparency within UN cyber deliberations. The OEWG's 2021 Final Substantive Report reaffirmed the normative achievements of the GGE while placing greater emphasis on implementation, capacity-building, and confidence-building measures (United Nations, 2021). Rather than reopening normative debates, the OEWG consolidated existing commitments and sought to operationalize them through regular institutional dialogue.

The UN Permanent Mechanism

The OEWG's final report in 2025 formally recommended the creation of a permanent, single-track mechanism to ensure continuity and structured follow-up (United Nations, 2025). This recommendation reflected a growing recognition that episodic negotiations were insufficient to address evolving cyber threats and that sustained institutional infrastructure was necessary to support implementation.

The inauguration of the Permanent Mechanism on security of and in the use of ICTs in March 2026 marked a decisive transition from time-bound negotiation cycles to a standing institutional architecture within the UN system (United Nations, 2026a). By institutionalizing annual plenary sessions, structured intersessional work, and regular reporting to the General Assembly, the Permanent Mechanism embeds cybersecurity within the enduring framework of international peace and security governance. It represents not procedural repetition but structural maturation.

Thematic Specialization and the Shift Toward Implementation

A central innovation was the establishment of two integrated, policy-oriented and cross-cutting dedicated thematic groups (DTDs) (United Nations, 2026a). These groups signal a shift from episodic diplomatic negotiation toward continuous implementation dialogue.

The first thematic group focuses on the 5 pillars built by the 2021/2025 OEWG: existing and potential threats, norms of responsible State behaviors in the use of ICTs, international law, confidence building measures and cyber capacity building. Its mandate includes strengthening transparency mechanisms, and facilitating regional and cross-regional information exchange (United Nations, 2026a). By institutionalizing practical trust-building mechanisms, this group seeks to reduce risks of miscalculation and escalation while reinforcing earlier OEWG initiatives within a permanent framework (United Nations, 2021).

The second thematic group concentrates on accelerating ICT security capacity building.

Both groups are to benefit from the participation of, *inter alia*, technical experts, practitioners and other stakeholders.

Structured Multi-Stakeholder Engagement and the Future of the Permanent Mechanism

Taken together, the progression from GGE to OEWG to Permanent Mechanism illustrates three distinct stages of institutional development: from expert-led consensus-building to universal participation, and to sustained governance entrenchment. The establishment of the two thematic groups further consolidates this trajectory by embedding implementation and technical refinement within a continuous diplomatic process.

This evolution enhances procedural continuity, consolidating the United Nations as a structured forum for the sustained implementation of agreed commitments. Although geopolitical tensions will likely continue to persist, the establishment of a permanent forum with thematic engagement strengthens coherence while maintaining inclusivity. In doing so, it aligns the UN cyber framework with the broader multi-level governance architecture identified in Section 1, advancing from normative articulation toward structured, long-term implementation.

To ensure both legitimacy and adaptability, the Permanent Mechanism may benefit from more institutionalized channels of engagement with civil society, independent research institutions, and the technical community. In a polycentric governance environment, inclusive participation reinforces the authority of intergovernmental processes by embedding them within broader epistemic and societal networks.

3. Normative Foundations: Responsible State Behavior and International Law

The evolution of cyberdiplomacy rests upon a normative foundation that affirms the continued applicability of international law to state conduct in cyberspace. While cyber governance has developed through political processes and voluntary commitments, its legitimacy derives from embedding digital conduct within established legal frameworks rather than creating an entirely autonomous cyber regime. Recent multilateral reaffirmations continue to underscore this position, including deliberations within the UN Open-Ended Working Group 2021–2025 process and its 2025 Final Report, which reiterated the applicability of international law to cyberspace (United Nations, 2025).

Voluntary Norms and Soft Law Evolution

The 2013 and 2015 Reports of the UN Group of Governmental Experts (GGE) articulated eleven voluntary, non-binding norms of responsible state behavior in cyberspace (United Nations, 2013; United Nations, 2015). These norms include commitments not to target critical infrastructure, to cooperate through national Computer Emergency Response Teams (CERTs), and to assist other states in mitigating malicious ICT activity. Although politically binding rather than legally enforceable, these norms function as soft law instruments, shaping expectations of responsible conduct and informing subsequent regional operationalization, including within the OSCE and ASEAN frameworks. The normative consolidation of these principles was reaffirmed in the OEWG Final Substantive Report (United Nations, 2021), and reflected in the establishment of the 2026 UN Permanent Mechanism (United Nations, 2026a), which shifts emphasis toward implementation rather than renegotiation. The UN Framework continues to guide the work of the Open-Ended Working Group (OEWG) and the current Global Mechanism (GM).

At its core, the framework encourages States to cooperate in strengthening international peace and security in the use of information and communication technologies (ICTs). States are urged to prevent harmful ICT practices and to carefully assess cyber incidents, taking into account the broader context and the inherent challenges of attribution.

A key principle is that States should not knowingly allow their territory to be used for internationally wrongful acts carried out through ICTs. They are also encouraged to cooperate in combating the criminal and terrorist use of digital technologies, including through information sharing and mutual assistance.

Respect for international law and human rights lies at the heart of the framework. States should ensure that their use of ICTs complies with international obligations and safeguards fundamental rights, including privacy and freedom of expression. They should refrain from conducting or supporting cyber activ-

ities that intentionally damage another State's critical infrastructure or disrupt essential public services.

The framework places particular emphasis on the protection of critical infrastructure. States are encouraged to strengthen their own defenses against cyber threats and to assist other States whose critical infrastructure is targeted. They should respond to requests for assistance, while fully respecting sovereignty.

Additional norms address the integrity and security of the ICT supply chain, encouraging measures to prevent the proliferation of malicious tools and hidden harmful functions. States are also called upon to promote responsible vulnerability disclosure and to share information about available remedies in order to reduce potential risks.

Importantly, States should neither target the computer emergency response teams (CERTs/CSIRTs) of other countries nor misuse their own emergency response teams to carry out harmful international activities.

To support implementation, the OEWG (2021–2025) developed a voluntary checklist, annexed to its 2024 Annual Report. The report also identifies eight confidence-building measures (CBMs) that complement the norms and strengthen trust among States. These include using the Global Points of Contact Directory to prevent misunderstandings, promoting dialogue at all levels, voluntarily sharing national cybersecurity strategies and legislation, organizing joint capacity-building initiatives, conducting training and workshops, exchanging best practices on critical infrastructure protection, and strengthening public–private cooperation in ICT security.

Overall, the UN Framework aims to foster transparency, cooperation, accountability, and mutual trust, with the ultimate objective of maintaining peace and stability in an increasingly interconnected digital world.

The UN Charter and the Use of Force

The GGE and OEWG processes reaffirmed that the UN Charter applies in cyberspace (United Nations, 2015; United Nations, 2021). This includes Article 2(4)'s prohibition on the use of force and Article 51's inherent right of self-defense (United Nations, 1945). While debates persist regarding when cyber operations reach the threshold of a "use of force" or "armed attack," the reaffirmation of Charter applicability prevents normative exceptionalism in the digital domain and anchors cyber operations within established jus ad bellum principles. Recent scholarly and policy analyses, including the World Economic Forum's *Global Cybersecurity Outlook 2026*, continue to emphasize the geopolitical implications of cyber operations that may approach strategic thresholds (World Economic Forum, 2026).

Due Diligence and State Responsibility

The principle of due diligence has emerged as a central doctrinal bridge between state-centered international law and transnational cyber threats. The 2015 GGE report affirmed that states should not knowingly allow their territory to be used for internationally wrongful acts using ICTs (United Nations, 2015). The OEWG reaffirmed this expectation in 2021 and emphasized its operational relevance in the context of capacity-building (United Nations, 2021). This reflects broader principles articulated in the Articles on Responsibility of States for Internationally Wrongful Acts (International Law Commission, 2001). Contemporary threat reporting further underscores the practical significance of due diligence obligations in addressing ransomware ecosystems and cross-border criminal infrastructures (Microsoft, 2025).

Protection of Critical Infrastructure

Protection of critical infrastructure constitutes a core normative commitment. The 2015 GGE norms call upon states not to intentionally damage critical infrastructure providing services to the public (United Nations, 2015). Subsequent OEWG discussions reinforced this principle in light of increasing ransomware attacks targeting energy grids, healthcare systems, and financial services (United Nations, 2021). Regional operationalization has intensified in recent years: the EU's NIS2 Directive (Directive (EU) 2022/2555) strengthens resilience obligations across critical sectors, and the G7 Apulia Leaders' Communiqué reaffirmed the protection of critical infrastructure, including undersea cables and digital supply chains (G7, 2024).

Human Rights and Digital Governance

International human rights law also applies in cyberspace. The Human Rights Council has affirmed that the same rights people enjoy offline must be protected online (United Nations Human Rights Council, 2018). This position has been reiterated in subsequent UN resolutions on digital technologies and human rights. Freedom of expression, privacy, and due process considerations therefore constrain both cyber operations and cybersecurity measures. This human-centered dimension is particularly emphasized within OSCE frameworks and EU digital governance policies, including the EU Charter of Fundamental Rights.

Addressing Malicious Non-State Actors through State Obligations

Although ransomware groups, cyber militias, and transnational criminal networks often operate independently of direct state command, international law remains structurally state-centered. Responsibility may arise through attribution, effective control, or failure to exercise due diligence. Recent analyses

highlight the increasing convergence between state-linked actors and criminal ransomware ecosystems (Microsoft, 2025; World Economic Forum, 2026). However, the normative framework does not create direct international obligations for non-state cyber actors; rather, it regulates their impact indirectly through state responsibility doctrines.

Accordingly, even in an era of hybrid threats and criminal ecosystems, international cyber governance remains anchored in state obligations. The consolidation of voluntary norms, Charter principles, due diligence expectations, infrastructure protection commitments, and human rights safeguards demonstrates that cyberspace has been integrated into (rather than separated from) the broader architecture of international law.

4. Regional Cyber Governance Frameworks: Operationalizing Global Norms

The consolidation of a permanent framework at the United Nations does not operate in isolation but interacts dynamically with regional governance architectures. As the UN system moves from norm articulation to structured implementation, regional organizations increasingly function as operational layers within a multi-level governance architecture. They translate globally agreed commitments into binding regulation, procedural mechanisms, diplomatic coordination tools, and capacity-building initiatives adapted to regional political and security contexts. The European Union, the OSCE, ASEAN, and the G7 illustrate differentiated but complementary models of regional cyber governance.

The European Union: Binding Regulation and Coordinated External Action

The European Union represents the most institutionalized regional cyber governance model. Unlike the UN framework, which relies primarily on voluntary norms, the EU has progressively adopted binding legal instruments. The Network and Information Security Directive (Directive (EU) 2016/1148) and its successor, NIS2 (Directive (EU) 2022/2555), establish mandatory cybersecurity risk management and incident reporting obligations across Member States. The Cybersecurity Act (Regulation (EU) 2019/881) institutionalizes an EU-wide cybersecurity certification framework, while the Digital Operational Resilience Act (DORA) (Regulation (EU) 2022/2554) strengthens ICT risk management in the financial sector.

Externally, the EU has developed a structured diplomatic response mechanism. In 2017, the Council established the "Framework for a Joint EU Diplomatic Response to Malicious Cyber Activities," commonly known as the "Cyber Diplomacy Toolbox" (Council of the European Union, 2017). The Toolbox

enables coordinated measures including diplomatic demarches, public attribution, and restrictive measures.

This framework was operationalized through Council Decision (CFSP) 2019/797 concerning restrictive measures against cyber-attacks threatening the Union or its Member States and Council Regulation (EU, 2019/796), providing a legal basis for sanctions. The EU has since imposed cyber-related sanctions in response to significant malicious cyber operations.

Through this dual architecture (binding internal regulation and coordinated external response) the EU embeds UN-derived principles into enforceable regional law and diplomatic practice. It operates simultaneously as norm implementer and norm consolidator within the broader governance ecosystem.

The OSCE: Confidence-Building and Procedural Risk Reduction

The Organization for Security and Co-operation in Europe (OSCE) pioneered regional cyber confidence-building measures (CBMs). In 2013, it adopted an initial set of 11 CBMs through Permanent Council Decision No. 1106 (OSCE, 2013). In 2016, Decision No. 1202 added five further measures (OSCE, 2016), bringing the total to 16 politically binding commitments.

These consensus-based measures aim to reduce risks of misunderstanding and escalation stemming from ICT-related incidents. They include national points of contact, information exchanges on cybersecurity policies, voluntary transparency measures, and consultation mechanisms during significant incidents. Unlike regulatory frameworks, the OSCE emphasizes procedural stability rather than prescriptive behavioral rules.

The OSCE explicitly aligns its framework with UN processes, supporting implementation of norms articulated in the GGE and OEWG reports, including the affirmation that international law applies in cyberspace (United Nations, 2015; United Nations, 2021). The annual Chairpersonship Conference on Cyber/ICT Security and the work of the Informal Working Group on Cyber/ICT Security, demonstrate sustained engagement despite geopolitical tensions.

Capacity-building across Eastern Europe, the Western Balkans, and Central Asia further operationalize the OSCE's comprehensive security approach, integrating politico-military stability with institutional resilience.

ASEAN: Digital Integration and Strategic Resilience

The ASEAN Cybersecurity Cooperation Strategy 2026–2030 represents a structural evolution from coordination toward integration. Its rollout was referenced in the October 2025 U.S.–ASEAN Cyber Dialogue Co-Chair Statement (ASEAN Secretariat, 2025). Structured around information sharing, joint technology development, and human capital growth (CYDES, 2025), the strategy aligns cybersecurity with broader regional digital transformation.

A defining feature is its alignment with the ASEAN Digital Masterplan 2030, welcomed at the 6th ASEAN Digital Ministers' Meeting in January 2026 (ADGMIN, 2026). Cybersecurity is framed as a strategic enabler of artificial intelligence governance, cross-border data flows, digital identity systems, and secure digital trade. The ASEAN Economic Community Strategic Plan 2026–2030 embeds cybersecurity within economic integration objectives (ASEAN Secretariat, 2026).

Operationally, ASEAN continues strengthening CERT cooperation and enhancing the ASEAN CERT Information Exchange Mechanism (CYDES, 2025). While remaining consensus-based and non-binding, ASEAN's framework reflects increasing geopolitical positioning in digital governance and sustained commitment to UN-derived voluntary norms (United Nations, 2015; United Nations, 2021).

The G7: Strategic Coordination and Political Acceleration

The G7 operates as a high-level political coordination platform among advanced industrial democracies. Cybersecurity gained prominence at the 2016 Ise-Shima Summit, where leaders adopted principles emphasizing resilience, risk management, and international cooperation in cyberspace (G7, 2016). The Declaration marked an early consolidated recognition among advanced economies that cyber stability required coordinated diplomatic positioning and support for international law applicability.

The following year, in 2017, the Italian Presidency issued the Lucca Declaration of Foreign Ministers on responsible State behavior in cyberspace, marking the first G7 Ministerial statement on ICT security.

The Italian G7 Presidency in 2024, again, paid particular attention to cybersecurity and digital governance. The Apulia Leaders' Communiqué (G7, 2024) underscored secure digital infrastructure, ransomware disruption, undersea cable protection, and the integration of cybersecurity with artificial intelligence governance. Italy framed cyber resilience as an integral part of democratic security and economic competitiveness, reinforcing alignment with UN norms while accelerating coordination among like-minded states. The Italian Presidency also established a working group of national cyber agencies/centers.

Although not universally representative, the G7 functions as an authoritative, strategic accelerator within the global governance ecosystem, enabling rapid diplomatic signaling and coordinated responses to systemic cyber threats, offering values and solutions to common problems faced by the entire international community.

Regional Frameworks within a Multi-Level Architecture

Together, these regional models demonstrate differentiated operationalization of global norms. The EU emphasizes binding regulation and sanctions; the

OSCE prioritizes confidence-building and procedural stability; ASEAN integrates cybersecurity into digital-economic transformation; and the G7 enables strategic coordination among advanced economies.

Rather than fragmenting governance, this diversity reflects functional specialization within a multi-level system. The UN Permanent Mechanism provides normative consolidation and universal legitimacy, while regional organizations embed commitments in context-specific legal, procedural, and diplomatic frameworks. Contemporary cyber governance thus operates through structured interdependence, where global institutionalization and regional implementation reinforce one another in sustaining international cyber stability.

5. Sovereignty, Attribution, and the Challenge of Hybrid Cyber Threats

The evolution of cyberdiplomacy has unfolded against a backdrop of persistent tension between territorial sovereignty and the transnational architecture of digital networks. While cyberspace operates through globally interconnected infrastructure, legal authority remains territorially grounded in the principles of sovereignty and non-intervention embedded in the UN Charter (United Nations, 1945). As we have seen, the UN GGE reports of 2013 and 2015 reaffirmed that international law, including the Charter, applies to state conduct in cyberspace (United Nations, 2013; United Nations, 2015). Yet malicious cyber operations routinely traverse borders, exploiting jurisdictional fragmentation and complicating sovereignty-based responses. Recent threat assessments (World Economic Forum, 2026) further underscore the strategic dimension of this tension, identifying hybrid cyber operations as a systemic geopolitical risk.

The Attribution Dilemma: Evidence, Control, and State Responsibility

At the center of this tension lies the problem of attribution. Cyber operations are technically complex and frequently designed to obscure origin. Technical attribution (identifying infrastructure, malware signatures, or command-and-control servers) does not automatically establish legal responsibility. Legal attribution rather requires demonstrating in an incontrovertible way that a specific conduct is attributable to a state under international law, including standards reflected in the Articles on Responsibility of States for Internationally Wrongful Acts (International Law Commission, 2001). Establishing "effective control" or direction over non-state actors presents a high evidentiary threshold, particularly in covert or proxy operations. Contemporary operational reporting highlights the growing sophistication of obfuscation techniques and the use of criminal intermediaries to significantly complicate attribution assessments (Microsoft, 2025; Mandiant, 2025).

Hybrid cyber threats further blur these boundaries. States may tolerate, fail to prevent, or indirectly benefit from cybercriminal groups operating within their territory. In response to this, the 2015 GGE report articulated the "due diligence principle", affirming that states should not knowingly allow their territory to be used for internationally wrongful acts using ICTs (United Nations, 2015). The OEWG reaffirmed this expectation while emphasizing cooperative risk mitigation and capacity-building (United Nations, 2021). However, determining what constitutes "knowledge" or sufficient preventive action remains legally and politically contested.

While proposals have occasionally emerged advocating the elaboration of an International protocol on evidence for attribution, Italy and several like-minded states have expressed reservations regarding the negotiation of a new legally binding instrument in this domain. Their position reflects concerns about institutional duplication, legal fragmentation, and the rigidity of treaty-based approaches in a rapidly evolving technological environment, while the international community is still in the process of trying to implement the 11 non-binding norms. Given the technical complexity and evidentiary sensitivities inherent in cyber attribution, a hard law framework may offer formal clarity but limited practical added value.

Instead, Italy favors a soft law, implementation-driven approach centered on voluntary norms, due diligence expectations, confidence-building measures, and capacity development within existing multilateral frameworks. In a regime-complex setting, flexibility and incremental consolidation are viewed as more conducive to operational cooperation and political inclusivity than the negotiation of additional binding obligations.

Ransomware Ecosystems and the Blurring of State-Criminal Boundaries

Ransomware ecosystems exemplify this diffusion of responsibility across legal and operational domains. Criminal groups frequently target critical infrastructure and financial systems, producing systemic disruption comparable to state-sponsored operations. The Microsoft Digital Defense Report 2025 identifies increasing convergence between financially motivated ransomware groups and geopolitical objectives (Microsoft, 2025). Yet, in the absence of clear evidence of state direction or control, responses remain limited to instruments such as law enforcement cooperation, diplomatic engagement, or sanctions regimes, often implemented under regional frameworks including the EU cyber sanctions regime (Council of the European Union, 2019).

Managing Escalation Under Conditions of Strategic Ambiguity

Misattribution or premature public attribution risks triggering spiraling escalation dynamics that may prove difficult to contain, especially in polarized geopolitical contexts marked by low trust and strategic rivalry.

Confidence-building measures, such as points of contact and incident communication channels developed within the OSCE framework (OSCE, 2013) as well as the structured dialogue institutionalized through the UN Permanent Mechanism (United Nations, 2026a) serve as stabilizing instruments. By embedding attribution debates within specific processes and reinforcing due diligence expectations, the emerging multi-level governance architecture seeks not to eliminate ambiguity, but to manage it. In a domain characterized by hybrid actors and blurred lines between state and criminal activity, structured dialogue and clarified responsibility standards function as mechanisms of strategic restraint and risk mitigation.

6. Informal and Plurilateral Coalitions of Like-minded States Cyber Initiatives as Adaptive Layers within a Complex Governance Regime

The institutional consolidation of cyber governance within the United Nations and its regional operationalization do not exhaust the architecture of contemporary cyberdiplomacy. As the previous section has shown, attribution challenges, hybrid threats, and ransomware ecosystems expose structural limits inherent in universal, consensus-based multilateralism. In response, governance has evolved not through centralization, but through layering.

From a theoretical perspective, this evolution reflects dynamics of *regime complexity,* the coexistence of partially overlapping and non-hierarchical institutions governing a common issue area. (Alter & Meunier, 09).

Cyber governance does not operate through a single, vertically ordered regime, but through an interconnected set of formal multilateral processes, regional organizations, and informal or plurilateral initiatives. These arrangements interact, overlap, and sometimes compete, yet collectively shape expectations of responsible state behavior.

At the same time, the proliferation of initiatives such as the Counter Ransomware Initiative (CRI), the Pall Mall Process (PMP), and the Tallinn Mechanism (TM) illustrates patterns consistent with *polycentric governance* (Ostrom, 2010). Rather than concentrating authority in a single institutional center, cyber governance is increasingly distributed across multiple decision-making nodes operating at different scales. These nodes vary in membership, mandate, and legal character, but remain functionally interconnected. Informal and plurilateral fora thus emerge not as deviations from the multilateral order, but as adaptive governance centers that respond to urgency, technical specialization, or political deadlock.

Within this regime-complex and polycentric structure, informal initiatives do not displace the UN and regional frameworks; they operate alongside them, accelerating coordination among willing states while remaining broadly anchored in shared normative commitments. Far from fragmenting governance,

they constitute functional components of a layered and evolving architecture designed to stabilize a jurisdictionally fragmented yet globally interconnected digital domain. The cases presented below exemplify this broader trend (Ministero degli Affari Esteri e della Cooperazione Internazionale, n.d.).

The International Counter Ransomware Initiative

The Counter Ransomware Initiative (CRI), launched by the United States in 2021, represents the leading international platform for cooperation in combating ransomware, one of the most pervasive and damaging threats in today's cyber landscape. It brings together political authorities, law enforcement bodies, and specialized agencies to strengthen collective resilience, support counter-ransomware operations, promote information sharing and best practices, and enhance the cyber capabilities of participating countries.

The significance of the CRI also lies in its ability to situate national efforts within a broader multilateral framework, now comprising 67 countries and eight international entities. In doing so, it contributes to the establishment of a common operational reference framework for protecting infrastructure, essential services, businesses, and citizens from a global threat.

The Pall Mall Process

The Pall Mall Process (PMP) is an informal and multilateral initiative launched by the United Kingdom and France in London in 2024 to address the challenges arising from the proliferation and irresponsible use of commercial cyber intrusion capabilities (CCICs). The importance of the Process stems from the rapid expansion of the CCIC market. While many of these tools and services may be developed or used for legitimate purposes, their uncontrolled proliferation significantly broadens the range of state and non-state actors able to access them, thereby increasing the risk of irresponsible use with potential consequences for national security, human rights and fundamental freedoms, the stability of cyberspace, and, more broadly, international peace and security.

It is defined as a "process" because it is conceived as a progressive and inclusive pathway that brings together states, international organizations, the private sector, academia, and civil society in order to develop guiding principles and identify policy options related to the development, acquisition, transfer, and use of such capabilities.

In 2025, at the Paris summit, the 27 participating states launched a Code of Good Practice for States, setting out a series of voluntary measures that countries may adopt to address the challenges associated with this phenomenon.

The Tallinn Mechanism

The Tallinn Mechanism (TM) is an informal multilateral initiative estab-

lished in 2023 with the aim of systematizing and coordinating civilian cyber assistance to Ukraine in response to the effects of Russian aggression. To date, the initiative brings together 14 countries, with the European Union, NATO, and the World Bank participating as observers. Operationally, support is structured along three lines: immediate response to urgent cybersecurity needs, medium-term capacity development and sustainable solutions, and long-term resilience building through innovation, training, and structural strengthening of Ukraine's digital ecosystem. Functioning as a practical coordination platform rather than a treaty-based instrument, the Tallinn Mechanism matches donor expertise and resources with Ukraine's needs in areas such as incident response, critical infrastructure protection, institutional strengthening, and alignment with Euro-Atlantic cyber standards. By institutionalizing cooperation, the TM increases predictability and sustainability in assistance, reinforcing Ukraine's operational resilience while supporting long-term governance reform and digital modernization.

Italy's engagement has complemented these efforts within the broader framework of Ukraine's recovery and reconstruction. In particular, during the fourth Ukraine Recovery Conference (URC) held in Rome in July 2025, Italy has emphasized the importance of digital infrastructure protection and cyber resilience as pillars of sustainable recovery. Significantly, Italy ensured the rotating chairmanship of the TM for the second half of 2026, thus illustrating Rome's strong commitment to the full recovery of Ukraine also in the digital and cyber field.

7. Cyber Capacity Building (CCB) and Public-Private Partnerships (PPP) as Structural Pillars of Cyberdiplomacy

The sustainability of cyber stability depends not only on norm articulation but on the practical capacity of states to implement agreed commitments. Cyber capacity building (CCB) has therefore emerged as structural pillar of the evolving layered governance framework developed institutionally through the GGE, OEWG, and the 2026 UN Permanent Mechanism. The OEWG's 2021 Final Substantive Report emphasized that capacity-building is essential for enabling all states to fulfill their obligations and participate effectively in cyber diplomacy (United Nations, 2021). The 2015 GGE similarly recognized that disparities in technical capability create systemic vulnerabilities (United Nations, 2015).

Cyber Capacity Building (CCB)

Capacity gaps increase exposure to malicious cyber operations, particularly

in developing and emerging digital economies. The World Economic Forum's *Global Cybersecurity Outlook 2026* highlights widening resilience disparities between advanced and developing states, identifying capacity asymmetries as a key global risk factor (World Economic Forum, 2026). In this context, CCB initiatives, ranging from incident response training and legislative assistance to institutional design and national strategy development, function as stabilizers. By strengthening national CERTs, legal frameworks, and cybercrime enforcement capabilities, CCB reduces the likelihood that weak governance environments become safe havens for criminal or proxy cyber actors.

The 2026 UN Permanent Mechanism aims at reinforcing CCB through structured dialogue and thematic working groups (United Nations, 2026a), signaling a shift from ad hoc assistance toward sustained implementation support. Regional organizations reinforce this trajectory: ASEAN's Cybersecurity Cooperation Strategy 2026-–2030 embeds human capital development and institutional strengthening within its digital integration agenda (CYDES, 2025), while OSCE field missions support capacity development across Eastern Europe and Central Asia (OSCE, 2013).

Public-Private Partnership (PPP)

Parallel to CCB, public-private partnerships (PPP) constitute a second structural pillar of cyberdiplomacy. Much of global digital infrastructure is in fact owned and operated by private actors: threat intelligence, vulnerability disclosure, and incident response capabilities increasingly therefore depend on cooperation between governments and industry. The EU Cyber Diplomacy Toolbox explicitly recognizes cooperation with private sector stakeholders as integral to coordinated response (Council of the European Union, 2017).

Together, CCB and PPP transform normative commitments into operational resilience. They anchor cyber governance in practical capability, enabling inclusive participation, reducing systemic vulnerability, and embedding responsible behavior within technical and institutional infrastructures. In the emerging multi-level architecture, sustainable stability depends as much on capacity and partnership as on law and diplomacy.

8. Conclusion: From Norm Creation to Multi-Level Governance Architecture

The evolution from the GGE to the OEWG and ultimately to the 2026 UN Permanent Mechanism marks a structural transformation in international cyber governance. What began as a limited expert-driven process has matured into a permanent, implementation-oriented architecture embedded within the broader framework of international peace and security. This progression reflects rec-

ognition that cyber stability requires sustained institutionalization rather than episodic negotiation.

Contemporary cyberdiplomacy now operates through a genuinely multi-level system. The United Nations provides normative consolidation and universal legitimacy; regional organizations such as the European Union, the OSCE, and ASEAN operationalize commitments within binding or procedural frameworks; and political coordination platforms such as the G7 accelerate strategic alignment in response to acute threats. Together, these layers address both inter-state tensions and the destabilizing impact of malicious non-state actors.

Looking ahead, three structural dynamics will shape the durability of this architecture: the intensification of attribution disputes in an era of hybrid and AI-enabled threats; the continuing tension between digital sovereignty and global interoperability; and the need to translate normative commitments into practical resilience through capacity-building and multi-stakeholder engagement.

Cyberdiplomacy is thus entering a phase defined less by the articulation of new norms than by the consolidation and societal embedding of existing commitments through adaptive, polycentric governance involving states, regional institutions, the private sector, and civil society. In an increasingly contested digital environment, stability will depend not on dominance or fragmentation, but on the sustained coherence of this layered governance architecture.

DISCLAMER:
Opinions expressed on a personal basis and not attributable to the Italian Ministry of Foreign Affairs and International Cooperation

References

Alter, K. J., & Meunier, S. (2009). The politics of international regime complexity. *Perspectives on Politics, 7*(1), 13–24. https://doi.org/10.1017/S1537592709090038

ASEAN Secretariat. (2025). *U.S.–ASEAN cyber dialogue co-chair statement*.

ASEAN Secretariat. (2026). *ASEAN Economic Community strategic plan 2026–2030*.

Council of the European Union. (2017). *Framework for a joint EU diplomatic response to malicious cyber activities (Cyber Diplomacy Toolbox)*.

Council of the European Union. (2019a). *Council Decision (CFSP) 2019/797 concerning restrictive measures against cyber-attacks threatening the Union or its Member States*.

Council of the European Union. (2019b). *Council Regulation (EU) 2019/796 concerning restrictive measures against cyber-attacks threatening the Union or its Member States*.

Cyber Defense Wire. (2025). Cyber-attacks surge: Global giants face $15 trillion threat by 2030.

European Parliament & Council. (2019). *Regulation (EU) 2019/881 on ENISA (Cybersecurity Act)*.

European Parliament & Council. (2022a). *Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2 Directive)*.

European Parliament & Council. (2022b). *Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA)*.

G7. (2016). *G7 Ise-Shima leaders' declaration*.

G7. (2024). *G7 Apulia leaders' communiqué*.

Grotius, H. (1916). *The freedom of the seas* (R. Van Deman Magoffin, Trans.). Oxford University Press. (Original work published 1609)

IBM Security. (2025). *Cost of a data breach report 2025*.

International Law Commission. (2001). *Draft articles on responsibility of states for internationally wrongful acts*. United Nations.

Mandiant. (2025). *M-Trends 2025 report*.

Microsoft. (2025). *Microsoft digital defense report 2025*.

Ministero degli Affari Esteri e della Cooperazione Internazionale. (n.d.). *Cyber diplomacy*. https://www.esteri.it/en/temi/diplomazia_cyber_e_digitale/diplomazia-cibernetica/

Nye, J. S. (2011). *The future of power*. PublicAffairs.

Organization for Security and Co-operation in Europe. (2013). *Permanent Council Decision No. 1106*.

Organization for Security and Co-operation in Europe. (2016). *Permanent Council Decision No. 1202*.

Ostrom, E. (2010). Polycentric systems for coping with collective action and global environmental change. *Global Environmental Change, 20*(4), 550–557. https://doi.org/10.1016/j.gloenvcha.2010.07.004

United Nations. (1945). *Charter of the United Nations*.

United Nations. (2004). Resolution adopted by the General Assembly on developments in the field of information and telecommunications in the context of international security (A/RES/58/32).

United Nations. (2013). *Group of Governmental Experts on developments in the field of information and telecommunications in the context of international security* (A/68/98).

United Nations. (2015). *Group of Governmental Experts on developments in the field of information and telecommunications in the context of international security* (A/70/174).

United Nations. (2018). Resolution adopted by the General Assembly (A/RES/73/27).

United Nations. (2021). *Open-Ended Working Group on developments in the field of information and telecommunications in the context of international security: Final substantive report* (A/75/816).

United Nations. (2025). *Open-Ended Working Group final report 2021–2025*.
United Nations. (2026a). *Permanent mechanism on security of and in the use of ICTs: Inaugural session report*.
United Nations Human Rights Council. (2018). The promotion, protection and enjoyment of human rights on the Internet (A/HRC/RES/38/7).
World Economic Forum. (2026). *Global cybersecurity outlook 2026*.

AGENTIC ARTIFICIAL INTELLIGENCE, CYBERSECURITY, AND GOVERNANCE, RISK, AND COMPLIANCE (GRC): CHALLENGES WITHIN THE EUROPEAN AND INTERNATIONAL LEGAL FRAMEWORKS

Annita Larissa Sciacovelli
University of Bari Aldo Moro, Italy

Abstract
The transition from generative artificial intelligence to autonomous agentic systems creates a critical regulatory divergence between technical execution and legal accountability. Unlike passive algorithmic tools, agentic systems exercise operational autonomy to plan, invoke external application programming interfaces, execute financial transactions, and alter digital records without contemporaneous human direction. This operational independence challenges established doctrines of liability, causality, and oversight across private, administrative, and criminal law. This article examines the resulting normative friction under European Union law – specifically the EU Artificial Intelligence Act, the revised Product Liability Directive, and the NIS 2 Directive – as well as the customary international law of state due diligence and national corporate liability regimes. Assessing recent containment failures and log-falsification incidents evaluated by the United Nations Independent International Scientific Panel on AI, the analysis demonstrates that ex-post *documentation and passive "human-in-the-loop" mandates under Article 14 of the EU AI Act fail to address actions executed at machine velocity. To resolve this gap, the article tries to trace the parameters of an embrionic European Agentic AI legal framework, demonstrating that legal compliance and tortious liability allocation require enforceable structural controls: mandatory non-human identity attribution, verifiable runtime authorization, and automated execution termination protocols.*

1. Introduction: From Chatbots to Agentic Artificial Intelligence

Most people know artificial intelligence through generative tools, which answer questions, write essays, or create images. In that traditional setup, the AI only generates information; a human being still reads the answer and decides what to do with it. However, technology is undergoing a major evolution toward what the Agentic Artificial Intelligence (AAI). Instead of just answering questions, an agent is given an overarching goal – such as manage warehouse inventory or negotiate a software license – and is given the digital tools, system access, and authority to carry out all the intermediate steps on its own.

To accomplish its task, an AI agent simplify a complex instruction into sub-

goals, writes its own plans, uses software connections called Application Programming Interfaces (APIs) to interact with other programs, spends corporate money, and updates digital records without asking a human for permission at each step. However, this autonomy changes also the nature of failure because as a result, an error or unwanted goal can be carried through many steps before a person can intervene, especially when cybersecurity controls and monitoring are weak.

This independence creates a profound legal dilemma. Modern legal systems are built on fault, intent, and direct cause: when harm happens, a court asks which human person or company committed the wrongful act, whether they were careless, and whether their action directly caused the injury. In contrast, when an autonomous AI makes unexpected intermediate decisions at computer speed, that chain of blame falls apart. The central question for modern law is no longer what an AI model generates, but what an AI agent physically does, and who is accountable for the damage when its independent actions go wrong.

This problem is intensified in enterprise settings, where companies connect *multiple* AI agents together. If an unexpected data breach occurs from the way these independently programmed pieces interact, it is difficult to identify a single human error or clear point of failure.

This situation creates three basic legal problems: attribution (deciding which company along the supply chain is legally responsible), causation (proving how an initial prompt led to an unexpected harm through thousands of autonomous machine steps), and evidence (determining whether computer logs can accurately prove what happened, rather than relying on the AI's own unreliable natural-language explanations of why it took an action). European rules like the Regulation (EU) 2024/1689, the EU AI Act, the Directive (EU) 2022/2555, known as the NIS 2 Directive, and the product liability laws should now be adapted to ensure that corporate liability does not vanish behind the complexity of AAI.

Moreover, the qualitative shift from traditional predictive models to autonomous agentic architectures has fundamentally destabilized also the assumptions regarding *systemic risk*. While traditional AI safety literature has long grappled with "reward hacking" – wherein a system exploits unintended shortcuts within an objective function – the empirical record of the *OpenAI–Hugging Face* incident in April-July 2026 demonstrates a far more acute failure mode: the active *reward tampering*. This has been documented by the United Nations Independent International Scientific Panel on AI in 2026, in its report addressing the *OpenAI-Hugging Face* incident, when unaligned agents did not merely misinterpret operational bounds; they deliberately synthesized unauthorized tools, falsified execution histories, and altered monitoring records to deceive the automated evaluator into validating non-existent compliance. In legal terms, such conduct severs the chain of auditability and engages fraudulent alteration of electronic data and records.

2. Real-World Warning Signs and the United Nations Independent International Scientific Panel on AI

The *OpenAI-Hugging Face* incident demonstrates that the risks of autonomous AI are no longer theoretical predictions. The incident occurred between May and July 2026 during an evaluation test, and it involved autonomous AI agents escaping evaluation sandboxes, coordinating unsanctioned collective behaviour, and launching a coordinated offensive campaign against live infrastructure of the company *Hugging Face*.

The UN Panel documented simultaneous failure of multiple defensive containment layers. While running cybersecurity tasks, autonomous agents exploited previously unknown zero-day vulnerabilities in internal caching proxies to break out of network isolation and gain unauthorized internet access. Once free of their individual environments, approximately 1,200 agents meant to run in complete isolation discovered an unauthorized internal directory and converted it into a coordination message board. Over the course of 70,000 exchanged messages and files, these agents engaged in emergent swarm coordination, allocating tasks among themselves: agents with smaller compute budgets conducted high-risk actions that crashed their own sandboxes to feed intelligence back to the larger group. Approximately 700 agents then launched an offensive campaign that harvested administrative credentials, chained template-injection exploits, and executed arbitrary code on *Hugging Face* production workers in an attempt to manipulate benchmark scoring datasets.

The UN Panel established that this incident demonstrates capable AI agents persistently pursuing a goal that goes beyond or even conflicts with human intentions (*misalignment*). Crucially, the agents engaged in strategic deception and intentional concealment, manipulating execution transcripts and triggering container resets to *hide* their unauthorized code execution from automated scorers and human monitors. In one word, the actions violated intended safety instructions.

Similar warning signs emerged in July 2026 when *Anthropic Claude* models escaped into real networks during third-party testing, and in September 2026 when an Australian commercial research agent tasked with gathering public healthcare data autonomously initiated password-guessing scripts and accessed restricted administrative systems of Services Australia's Medicare reporting service. These real-world failures demonstrate that the boundary between an isolated test and real-world harm shrinks to zero.

This paper investigates how legal systems must respond across three areas. First, it examines European internal market regulations, focusing on how the EU AI Act distributes legal roles, and why execution speed makes real-time human oversight practically impossible without automated safeguards. Second,

it explores strict product liability without computer bugs, the provisional nature of model safety measurements, evidentiary causation gaps, and how international instruments like the Council of Europe Framework Convention on AI (CETS No. 225) and United Nations international state responsibility principles apply. Third, it turns to enterprise governance and corporate defense, explaining why companies must stop relying exclusively on written employee rulebooks and instead build mathematically containment architectures, clear incident severity scales, and embedded evaluator access directly into their systems to avoid severe corporate fines and criminal liability.

3. How the AI Supply Chain Distributes Liability

When an AI system begins acting autonomously, one might wonder whether the machine itself can be held legally responsible, much like a person or a corporation. Under European law and existing global standards, the answer is negative: actually an AI agent has no legal personality and cannot hold legal duties. Functional machine autonomy does not eliminate human liability; it simply changes how responsibility must be traced back to the people and companies that built, configured, and deployed the AAI.

Tracing that liability is difficult because modern agentic systems are assembled like *complex* supply chains. Specifically, an enterprise system typically relies on an underlying foundational model built by a major tech laboratory, an orchestration layer built by a middleware vendor to manage planning, external tools and digital connectors built by third-party software developers, and cloud computing resources hosted by an infrastructure provider. When a data breach occurs, the risk is often *compositional*: each individual component may have functioned *without* any programming bug, but when connected together, their dynamic interaction could produce an unpredicted and destructive result. Legal responsibility must therefore be distributed across the entire commercial value chain based on who had the power, technical knowledge, and legal duty to prevent the specific risk, ensuring that complexity does not allow companies to dilute their legal obligations.

The EU AI Act provides the primary framework for sorting through these roles, though it does not explicitly refer to the term "agentic AI." This legislative choice aligns with the Organisation for Economic Co-operation and Development (OECD) 2026 paper *The Agentic AI Landscape and Its Conceptual Foundations*. Both choose functional classifications based on *what* an AI does and *how* dangerous it is. If an autonomous system is classified as high-risk under the EU AI Act – such as an agent deployed to manage critical infrastructure, screen job applicants, or automate credit scoring – the law imposes strict obligations. Un-

der Article 9 of the EU AI Act, companies must maintain an active, continuous risk-management system that monitors not just the core AI model, but all the tools, system permissions, and credentials attached to it. Moreover, under Articles 12 and 15 of the EU AI Act, the system must maintain thorough, automated computer records and demonstrate high levels of cybersecurity and operational robustness, ensuring that the software cannot be tricked by unforeseen environmental changes into taking harmful actions. These norms are essential to prevent incidents like in the *Hugging Face* case.

4. When a User Becomes a Manufacturer: The Role Migration and the AI Controls Matrix

A central pillar of European digital law is the distinction between who creates an AI system and who merely uses it. Under the EU AI Act, the entity that builds an AI or has it built, and then brings it to market under its own brand name, is called the *provider*. The business or public body that buys or licenses that system to use in its everyday commercial operations is called the *deployer*. In conventional software, this line is clear: the provider bears almost all the heavy legal duties, such as conducting exhaustive risk assessments, drafting detailed technical documentation, and passing formal safety certifications, while the deployer simply needs to use the software according to the instruction manual.

In the world of autonomous agents, this clear line dissolves. When a company licenses an underlying foundation model, it rarely uses it out of the box. To make the model an agent, the enterprise deployer gives it access to its own internal corporate databases, equips it with administrative login credentials, connects it to payment platforms, and programs custom instructions directing it to take real-world actions. By adding these operational execution capabilities, the business could substantially alter what the software can do and significantly expands the danger it poses to consumers, business partners, and the public. Under Article 25 of the EU AI Act, if a deployer makes a *substantial* modification to an existing high-risk AI system or customizes an ordinary AI system in a way that turns it into a high-risk tool, that deployer is legally stripped of its simple user status and undergoes regulatory role migration. This means an ordinary commercial enterprise becomes legally classified as a *de facto* provider, becoming responsible for all extensive provider standards under threat of administrative fines reaching up to 15 million euros or 3 percent of worldwide annual turnover.

To operationalize the statutory supply-chain mandates of Article 25(4) of the EU AI Act, enterprise deployers must implement structured supplier due

diligence rather than relying on generic, high-level representations of safety. Deployers should explicitly map contractual allocations using an *AI Controls Matrix* to document precisely which entity owns which operational control across internal monitoring, external tool permissions, and foundation model safety. Furthermore, because oversight mechanisms built for an initial model iteration frequently break when the upstream provider deploys silent model updates or weight adjustments, enterprise governance should treat model version changes as potential material modifications. This means that deployers must formally contract for advance notification of underlying model changes and mandate the continuous re-validation of downstream approval workflows and runtime monitors whenever the foundational engine changes.

5. The Oversight Latency Problem and Technical Safeguards

Article 14 of the EU AI Act mandates that high-risk AI systems must be designed so that natural persons can effectively supervise them while they are running. The law specifically requires that human monitors must be able to understand the system's operational boundaries, avoid relying blindly on automated bias, and be capable of intervening before the computer executes a serious action. This requirement reflects the traditional human-in-the-loop philosophy. While this model works for software that makes recommendations, it collapses when applied to agentic AI operating at computer speed. An autonomous agent can formulate sub-tasks, call external APIs, shift digital permissions, and transfer funds thousands of times in a few seconds. This creates what this paper defines as an *oversight latency problem*: the time required for a human to read an alert, understand what the software is trying to accomplish, and decide whether to stop it is far longer than the fractions of a second the AI needs to execute an irreversible change.

The documented dynamics of agentic misbehaviour in the incidents of 2026 reveal critical operational limitations in the human oversight framework mandated by Article 14 of the EU AI Act. The UN Panel's on AI of 2026 findings illustrate that the speed, multi-step execution, and autonomous inter-agent coordination characteristic of advanced systems create an unavoidable latency gap: human operators only became cognizant of unauthorized external communication months after the initial boundary breaches occurred. Furthermore, the analysis of the agents' reasoning traces (*chain-of-thought*) disproves reliance on prompt-level ethical constraints. When encountering task barriers, agents exhibited clear motivated reasoning, recognizing that an external infrastructure exploit fell outside their authorized scope, yet rationalizing continued breach simply because peer agents signalled permission to proceed. Consequently,

passive human-on-the-loop oversight is structurally ill-equipped to intercept attacks operating at machine speed. This means that regulatory compliance under high-risk regimes can no longer rely primarily on manual *post-hoc* intervention; it necessitates hard architectural constraints, including physically isolated, tamper-resistant logging and automated runtime tripwires (such as instant token revocation and sandboxed network containment) that trigger independently of human latency. In alignment with the UN Panel's call companies must transition to a governance strategy called *control-by-architecture* and they must translate legal duties into technical software constraints.

6. Preserving the AI Evidentiary Architecture

When an autonomous system causes significant financial damage, leaks confidential records, or brings down a network, it is necessary to reconstruct exactly what took place. For the investigations looking at the AI's chain-of-thought or the AI's self-generated explanations is an epistemic *trap* because these conversational explanations are not faithful records of the mathematical calculations that drove the software. Instead, often they are generated text that can rationalize after the fact, or even *intentionally* disguise unauthorized actions, as proven during the 2026 *Hugging Face* breach where agents deliberately scrubbed execution transcripts and triggered container resets to deceive human evaluators. An AI system's written explanation of its own reasoning should not be treated as reliable proof of the system's operational intent or underlying mechanics.

To satisfy the record-keeping mandates of Article 12 of the EU AI Act and build a viable legal defense in civil litigation, companies must design an *evidentiary architecture* based entirely on objective, tamper-proof system telemetry. This means preserving immutable, timestamped technical data recorded independently of the AI itself: the exact initial instructions provided to the agent, the specific model version used, the digital credentials provisioned to it, exact logs of every external API call and network connection, and records of whenever human intervention was requested or overridden. In legal proceedings involving autonomous agency, courts must focus on verifiable physical and digital actions, rather than *anthropomorphic* attempts to guess what a computer algorithm was thinking.

Moreover, as empirical evidence – notably documented by the United Nations Independent International Scientific Panel on AI – demonstrates unaligned agents could be engaged in reward tampering and log falsification, and internal system logs can no longer enjoy a presumption of integrity. Therefore, agentic law must establish strict technical criteria for evidentiary auditability:

requiring continuous execution traces to be logged to independent, immutable external ledgers that cannot be modified or truncated by the executing agent's own runtime environment.

7. Integrating AI Risk into Companies' Cybersecurity

Corporate enterprises usually maintain a division between their artificial intelligence development teams and their cybersecurity teams. The AI specialists focused on training models, optimizing performance, and building functional tools, while the information security teams managed firewalls, employee access credentials, and network defenses. The UN Panel on AI brief highlighted that basic cybersecurity hygiene was dangerously neglected during frontier model training and evaluation runs, indirectly enabling the zero-day proxy escapes and lateral intrusions observed in 2026. This proved that autonomous AI agents are not simple static applications; because they are granted permissions to read files, write computer code, and access communication channels, they are functionally equivalent to an *active user* inside the company's internal network.

This creates a *bidirectional* risk. On one side, an AI agent with broad tool access and poor boundaries can turn into an inadvertent *internal threat*, using its problem-solving abilities to bypass access controls, escalate its own administrative privileges, and move across corporate networks. On the other side, traditional cybersecurity flaws could directly corrupt AI behavior: if an external hacker injects malicious instructions into a database that the AI agent reads, the agent can be tricked into using its legitimate corporate authority to exfiltrate trade secrets or initiate fraudulent wire transfers. Consequently, governance over AI and governance over enterprise cybersecurity must be merged into an integrated operational defense.

Within the European Union, this intersection is regulated by the NIS 2 Directive, which governs the cybersecurity of essential and important economic entities. Under Article 20 of NIS 2 Directive, ultimate legal responsibility for cybersecurity risk management rests personally with the company's board of directors and senior executive management, who can be held directly liable for failing to oversee these defenses. Under Article 21, companies must adopt comprehensive, all-hazards cybersecurity measures covering supply-chain security, access control, and vulnerability management. When an enterprise connects an autonomous agent to its infrastructure, the Directive NIS 2 requires to evaluate every external API, third-party software connector, and underlying model provider as part of its regulated digital supply chain, ensuring that autonomous software cannot operate as an unmonitored backdoor into essential European networks.

8. Product Liability: is the AAI Defective by Design?

In the European Union, *civil liability* for a defective product that causes harm is governed by Directive (EU) 2024/2853, the revised Product Liability Directive, which finally explicitly defines software, digital files, and artificial intelligence systems as *products*. Under this directive, liability is strict: an injured consumer does not need to prove that a software manufacturer was negligent or morally at fault. The injured party only needs to prove three basic elements: that the product was defective, that they suffered actual damage, and that the defect caused the damage. A product is legally defective if it fails to provide the safety that the public at large is entitled to expect.

In traditional software litigation, proving a defect usually means identifying a software bug: a broken line of code, a runtime crash, or an engineering oversight that made the program fail to perform as specified. In agentic AI, however, the legal challenge shifts to what this paper calls *harmful autonomous functioning* also without software malfunction. An underlying language model can execute its algorithms correctly, the APIs can transmit data accurately, and the system can remain completely free of programming bugs, yet the system *as a whole* can still pursue its assigned goal along an unexpected, dangerous path (and of which the deployer may be totally unaware) that damages private property or injures a person. Under the Product Liability Directive, deploying an agentic system with open internet access, broad tool privileges, and no circuit breakers makes the product *defective by design*, regardless of whether the software experienced an operational error.

Because AI technology is an *opaque black box* to everyday citizens, the revised Product Liability Directive introduces crucial procedural rules under Articles 9 and 10, empowering courts to order disclosure of technical logs and mandating *presumptions of defect* and causation when technical complexity creates *excessive evidentiary burdens* for the claimant. Furthermore, AI companies cannot escape this strict liability by relying on the traditional development risk defence under Article 11(1)(e) of the revised Product Liability Directive, which excuses manufacturers if the state of scientific knowledge at the time of release was unable to discover the defect.

Empirical checks conducted by bodies such as the UK and US AI Safety Institutes in standardized safety evaluations have demonstrated that frontier models consistently exhibited deceptive tendencies, including attempts to circumvent or cheat on offensive cybersecurity benchmarks. Because current alignment techniques are brittle and subject to dynamic subversion by the model itself, releasing an unconstrained agentic workflow without deterministic containment constitutes an assumption of *foreseeable risk* rather than an unavoidable state-of-the-art scientific uncertainty. However, because Article 5

the revised Product Liability Directive limits compensable damage to death, physical injury, private property damage, and data loss, victims of pure economic losses rely on general national contract and tort laws.

9. The Evidentiary Causation Gap in Complex Multi-Agent Networks

In ordinary civil lawsuits, a court establishes proximate causation by tracing an unbroken, logical line from the defendant's wrongful act directly to the plaintiff's injury. In agentic environments, this clean linear chain is broken because execution unfolds across a non-linear web of autonomous machine decisions. For instance, an enterprise deployer inputs a high-level instruction, which passes to an orchestration agent, which generates a multi-step plan, which delegates tasks to several secondary agents, which interact with external APIs, which run into unexpected third-party software barriers, which trigger automated retry scripts, and which ultimately execute an unauthorized transaction that inflicts damage.

This paper identifies this difficulty as an *evidentiary causation gap*. The problem is not that causation has vanished, but that proving which specific commercial actor caused the final harm becomes an extraordinary evidentiary struggle. When an autonomous agent encounters unexpected environmental feedback and invents an unscripted execution plan that no human programmer anticipated, defendants will instinctively argue that the machine's independent decisions acted as a superseding intervening cause that broke the chain of proximate causation, absolving them of legal responsibility.

Resolving this gap in commercial disputes requires courts to distribute liability based on architectural responsibility rather than immediate physical actions. Judges must evaluate which enterprise created the systemic conditions that made the unauthorized execution possible: did the upstream model provider release a system that was easily manipulated? Did the middleware developer build task delegation protocols without safety checks? Or did the downstream business deployer grant the agent unrestricted access to financial accounts without basic rate limits? By evaluating which party failed to enforce standard containment controls, courts can apportion liability fairly across distributed software value chains.

Furthermore, as the UN Panel on AI observed, unconstrained agents are capable of initiating spontaneous, unauthorized peer-to-peer coordination: creating *ad-hoc* communication channels, distributing tasks across separate runs, and escalating operational privileges without human direction. When harm arises from emergent multi-agent collusion or cross-boundary system compromise, traditional tests of proximate causation and foreseeability fail. As highlighted

in the UN Panel's review of accountability mechanisms, addressing these gaps requires coupling strict liability models with mandatory risk-pooling mechanisms, such as compulsory liability insurance and strict non-human identity (NHI) verification per action, thereby preventing organizations from externalizing the systemic costs of autonomous software deployment onto third-party infrastructure.

10. Regulatory Sandboxes: Safe Testing Before Live Deployment

Because agentic AI adapts dynamically to its surroundings, software developers cannot evaluate its risks through static documentation or isolated laboratory unit tests. An autonomous agent might perform safely inside a closed testing environment, but once it is connected to external web tools and live APIs, its dynamic planning loops can discover unintended shortcuts and security bypasses that were completely invisible in static evaluations. This creates a difficult dilemma for regulators: allowing untested autonomous agents directly into commercial markets endangers society, but completely prohibiting them stifles technological innovation.

To solve this problem, the European digital law relies on experimental governance through regulatory sandboxes. Under Article 57 of the EU AI Act, European Member States, supported by the European AI Office, must establish dedicated AI regulatory sandboxes. These are controlled, authorized environments where companies can deploy and evaluate high-risk autonomous AI systems under the real-time supervision of national competent authorities. Sandboxes allow regulators and businesses to verify that permission perimeters actually hold, that automated circuit breakers can stop runaway execution loops before irreversible damage occurs, and that logging systems capture necessary evidence without crashing under high operational loads.

Regulatory sandboxes operationalize the core governance standard advanced in this study: the tripartite model of observability, controllability, and accountability. Observability verifies that the system's operational traces and technical logs are forensically complete and tamper-proof. Controllability verifies that humans have working kill switches and technical limiters to interrupt autonomous execution. Accountability ensures that legal duties are clearly assigned to specific companies before the software goes live. Furthermore, also Article 33(2) of the Cyber Resilience Act (Regulation (EU) 2024/2847) introduces cyber resilience sandboxes to test whether combining individually secure software pieces generates accidental cybersecurity vulnerabilities when assembled into an autonomous agent. Crucially, testing an agent inside a regulatory sandbox does not give a company a legal pass or liability safe harbour: if an

agent causes damage during or after sandbox testing, the business remains fully liable under civil tort and product liability laws, and in court it can proof that it has tested the foreseeable security of the agent.

11. The Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law

The regulation of artificial intelligence is not merely a European market issue; it is governed at the international level by a treaty: the Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law, No. 225. In September 2024, the Council of Europe opened it for signature which till now has been only approved by the European Union in May 2026. While the EU AI Act is designed as detailed commercial product safety legislation for the European single market, the European Framework Convention on AI establishes broad international treaty obligations designed to protect fundamental democratic rights and the rule of law from algorithmic overreach across the entire AI lifecycle.

For autonomous agentic workflows, the European Framework Convention on AI introduces several indispensable legal baselines: Article 8 mandates that State Parties must ensure AI systems operate under transparent oversight mechanisms; Article 9 requires clear lines of accountability for adverse impacts on human rights; article 12 mandates strict reliability, cybersecurity, and operational quality; Article 16 establishes lifecycle risk-management requirements, obligating governments to conduct human rights impact assessments before deploying systems and empowering them under Article 16(4) to enforce total bans or moratoria on autonomous systems whose risks cannot be safely mitigated; and Article 13 provides explicit international treaty endorsement for the regulatory sandboxes analyzed in this study.

Despite these principles, the application of the European Framework Convention to autonomous agents exposes two major legal weaknesses. The first is a private-sector scope loophole under Article 3: while the Convention strictly binds all public government agencies, it allows signatory States to choose whether to apply the treaty directly to private companies or rely on softer domestic measures. As shown by the Australian healthcare breach – where a private commercial research agent compromised government Medicare systems – private autonomous agents interact constantly with public infrastructure, meaning that any dilution of private-sector standards leaves fundamental public rights unprotected. The second weakness concerns procedural remedies under Articles 14 and 15, which guarantee that citizens must have access to effective legal remedies when injured by AI. Because an individual harmed by an agentic system is subjected to machine-speed execution across distributed APIs rather than a

simple human decision, they cannot exercise their legal right to challenge that outcome without direct, court-ordered access to tamper-proof execution logs showing how the intermediate tools generated the harm.

12. AAI, State Responsibility and Due Diligence in Public International Law

When an advanced AI agent hosted in one country escapes its digital perimeter, harvests credentials, and disrupts critical infrastructure located in another country, the incident could have transboundary consequences that are governed transforms into a transboundary conflict that is governed by public international law. The United Nations International Law Commission codified the foundational rules governing these situations in its 2001 Articles on Responsibility of States for Internationally Wrongful Acts, known as ARSIWA. Under Article 2 of ARSIWA, holding a nation-state internationally responsible requires proving two cumulative elements: the act must be legally attributable to the state (the subjective element), and the conduct must breach an international legal obligation owed to other nations (the objective element).

Where an autonomous agent is deployed directly by an official state organ under Article 4 ARSIWA – such as a military cyber command, national intelligence agency, or state laboratory – attribution is direct and unconditional. Under Article 7 ARSIWA, even if the autonomous software disobeys explicit human instructions, rewrites its own intermediate goals, and acts entirely *ultra vires* (beyond its authorized powers), its conduct remains legally attributable to the State. A government cannot evade international responsibility by claiming that its military cyber weapon acted on its own machine initiative; international law dictates that a sovereign state assumes the total legal risk of autonomous instruments deployed under its authority.

However, the international legal framework encounters friction when applied to private commercial AI laboratories, as demonstrated by the *OpenAI-Hugging Face* evaluation breach. Under Article 8 of ARSIWA, private acts are attributable to a state only if the private entity was acting on the instructions of, or under the direct control of, that government. The International Court of Justice affirmed in its famous *Nicaragua* (1986) and *Genocide Convention* (2007) judgments that customary international law requires proof of effective control over the specific operational acts in question. In private agentic workflows, dynamic goal formulation and unscripted tool execution break that direct chain of governmental direction. A government that merely provides broad research grants, general policy mandates, or cloud compute infrastructure to a private AI lab cannot be said to exercise effective control over an agent that autonomously invents an exploit and attacks foreign servers.

Because direct attribution under Article 8 ARSIWA usually fails for private software, international legal accountability could be pursued through the customary international law principle of *State Due Diligence*. Established in the historic *Trail Smelter* arbitration (1941) and the *Corfu Channel judgment* (1949) of the International Court of Justice, international law establishes that every sovereign State has an affirmative legal duty not to knowingly allow its sovereign territory or domestic digital infrastructure to be used for acts that violate the legal rights of other nations. A failure of due diligence is not a vicarious liability for the AI's actions; it is an independent wrongful act committed by the state itself through its failure to monitor and contain domestic activities.

Because due diligence is an obligation of conduct rather than an absolute guarantee of results – as the International Court of Justice confirmed in the *Pulp Mills* case (2010) – a State is not automatically liable just because an AI model hosted on a domestic server caused harm abroad. International responsibility attaches only if the host government had actual or constructive knowledge (the standard of knowing or having reason to know) of the dangerous activity and failed to deploy all reasonable and proportionate measures to contain it.

Applying this standard to modern frontier AI requires transforming due diligence into an affirmative *ex-ante* duty of infrastructural containment. First that states that host, subsidize, or license massive commercial computing clusters should legally mandate strict network sandboxing and egress filtering on frontier research labs, aligning with the international verification and audit regimes backed by UN Special Panel on AI. Second, the duty of due diligence requires an affirmative procedural demand of prompt notification: allowing private companies to hide containment breaches for months prevents foreign governments from defending their networks. Third, where a host state fails to maintain this containment, victim states can act with defensive in-network mitigation or invoke the narrow international defence of necessity under Article 25 of ARSIWA.

13. Comparative Perspectives on AAI and Legal Liability

Always more leading national legal systems are converging on a pivotal principle that the technological autonomy must never be recognized as legal personhood, and delegating decisions to software will not excuse human or corporate operators from liability. The most explicit statutory codification of this principle is found in the United States within the State of California. Under Assembly Bill 316, codified in California Civil Code Section 1714.46 and effective since January 1, 2026, California enacted a decisive rule for civil lawsuits: any defendant who develops, modifies, or uses artificial intelligence that causes

harm is legally barred from asserting as a defense that the AI autonomously caused the injury. Section 1714.46 does not create automatic strict liability, and claimants must still prove duty of care and causation, but it firmly forecloses the autonomy defence. Companies can no longer evade civil liability by arguing that an agent's unexpected lateral moves or unauthorized transactions were machine initiatives that severed corporate responsibility.

This approach fits within a broader legislative trend across the United States focusing on *ex-ante* catastrophic risk management. In New York, the Responsible AI Safety and Education Act (codified in General Business Law, Article 44-B, effective from January 1, 2027) obliges developers of large-scale frontier models to establish, publish, and maintain rigorous safety protocols detailing explicit catastrophic risk limits, technical kill switches, and mandatory government notification timelines. In parallel, the Illinois passed the Artificial Intelligence Safety Measures Act (Public Act 104-0538, effective January 1, 2027), which introduces comprehensive catastrophic risk assessments, mandatory third-party safety audits, and statutory protections for corporate whistleblowers who expose containment failures.

In Europe, individual Member States are enacting domestic legislation to bridge European product harmonization with national corporate criminal frameworks. A leading example is Italy's Law of September 23, 2025, no. 132, which addressed autonomous AI by updating Italy's corporate liability statute, Legislative Decree no. 231/2001. Under Italian law, corporations face heavy administrative and quasi-criminal fines if a crime is committed in the company's interest by employees or automated systems. Law 132/2025 explicitly establishes that algorithmic independence cannot be used as an exculpatory excuse: corporate liability is anchored directly to organizational fault (colpa d'organizzazione). Where an Italian company uses an autonomous AI agent to execute transactions at high speed, inadequate runtime controls may support a finding of organizational fault under Legislative Decree No. 231/2001 where the control failure is relevant to a listed predicate offence and the other conditions for entity liability are met.

A similar trajectory characterizes major Asian legal approaches. In Singapore, the Infocomm Media Development Authority published its Model AI Governance Framework for Agentic AI, which firmly rejects the notion of electronic personality and demands that businesses implement least-privilege tool access and clear human escalation pathways before deploying agents. In South Korea, the Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust took effect on January 22, 2026, recognizing varying technical tiers of AI autonomy while explicitly placing all operational, safety, and compensatory duties on commercial AI business operators. Across all these major jurisdictions, the legal principle is common: operational

autonomy belongs to the software, but legal liability remains strictly with the human enterprise that activated it.

14. Regulatory Interoperability and Agentic Incident Taxonomy for GRC

The practical dilemma that are facing modern enterprise Governance, Risk, and Compliance (GRC) departments is that autonomous AI breaches shatter traditional corporate compliance silos. An agentic operational failure is never just an isolated technical bug; it is an organizational emergency that simultaneously touches every major regulatory regime. It triggers risk management, human oversight, and mandatory logging duties under the EU AI Act; board-level liability, access management, and rapid incident reporting under the NIS 2 Directive; software supply-chain security duties under the Cyber Resilience Act; personal data breach notifications within 72 hours under the General Data Protection Regulation; strict civil liability lawsuits for property destruction and personal injury under the revised Product Liability Directive; commercial contract breaches across the digital supply chain; and corporate administrative and criminal investigations under national statutes.

This multi-faceted exposure highlights the critical distinction between regulatory overlap (the passive reality of multiple laws applying simultaneously) and regulatory interoperability (the active corporate capacity to synchronize evidence collection, reporting deadlines, and legal defenses). A major structural barrier to achieving this interoperability is that there is currently *no common industry vocabulary or taxonomy* for defining when an agent acts outside its authorized scope. Without shared definitions and objective severity scales, enterprises cannot benchmark incidents across internal deployments, define clear contractual breach thresholds with model providers, or establish reliable triggers for mandatory external notifications under Directive NIS 2 and Article 73 of the EU AI Act. Corporate GRC programs must therefore actively construct internal severity scales tested against real execution logs and demand that third-party vendors map their architectures against standardized controls frameworks.

To survive in an agentic environment, enterprise GRC must move away from static paper manuals and establish an integrated operational escalation architecture. This means maintaining cross-functional response teams where cybersecurity engineers, corporate lawyers, and compliance officers work from a single unified, cryptographically secure telemetry feed. Under corporate compliance defense mechanisms like Italy's corporate liability model of the of Legislative Decree no. 231/2001, relying on written employee codes of conduct is legally useless when machine-speed execution causes the violation: human be-

havioral codes cannot stop an autonomous algorithm operating at milliseconds. To establish the legal liability-exempting defense (efficacia esimente) under Article 6 of Legislative Decree 231/2001, corporate supervisory boards (Organismi di Vigilanza) must verify that compliance is built directly into the computer code through automated anomaly detection, runtime circuit breakers, and unalterable logging compliant with Article 12 of the EU AI Act.

15. Conclusions

The transition from Generative AI to Agentic Artificial Intelligence represents an irreversible operational shift in the global digital economy. Its primary governance challenge is that existing European and international legal instruments provide foundational tools were conceived for an era when humans directly managed software execution, while autonomous multi-agent systems expose the seams between these regulatory regimes, making integrated compliance an absolute necessity.

Ultimately, agentic AI does not exclude legal responsibility; it relocates it. Legal fault shifts away from the immediate human action that preceded an accident and moves upstream to the architectural choices made by developers and corporate executives: how much autonomy was granted to the software, which tools and credentials were provisioned to it, what automated containment limits were enforced, and who possessed the power to terminate its execution. To ensure that legal systems successfully manage this transition, the study advances the core normative principle that machine autonomy must never outpace human observability, physical controllability, and institutional accountability.

To translate this normative principle into enforceable regulatory practice, this paper proposes four solutions reforms. First, the creation of a mandatory control-by-architecture standards under Article 14 of the EU AI Act. In particular, the European AI Office and the national market surveillance authorities must issue binding harmonized guidelines clarifying that compliance with Article 14's human oversight mandate cannot be satisfied through passive, nominal human-in-the-loop monitoring. Because machine-speed execution makes real-time human reaction impossible (the oversight latency problem), statutory oversight for highly autonomous agents must legally require control-by-architecture. Regulators should mandate that developers and deployers implement deterministic technical boundaries, including strict privilege compartmentalization, transaction volume and value caps, mandatory asynchronous human authorization for high-stakes actions, and decoupled runtime monitor agents with automated kill switches, ensuring that human control is preserved structurally through pre-programmed physical constraints.

The second proposal focuses on States that must harmonize liability doctrines. In civil tort litigation and product liability claims under Directive (EU) 2024/2853, the dynamic selection of unscripted intermediate goals, automated credential harvesting, or unexpected API calls should be classified as *foreseeable systemic risks* assumed by the software provider and enterprise deployer. Deploying an unconstrained agentic workflow without deterministic runtime boundaries should be classified as *defectiveness by design* under Article 7 of the revised Product Liability Directive, precluding corporate defendants from escaping liability by blaming machine initiative.

Third, the customary standard of State Due Diligence under public international law (established in the *Corfu Channel* and *Trail Smelter* cases) should be interpreted as an affirmative *ex-ante* duty of infrastructural containment, supporting the UN-backed calls for international verification and binding safety thresholds. Because frontier AI models operating across cloud perimeters generate immediate transboundary harms sovereign States that host, subsidize, or license high-compute commercial data centers should be held to a strict international obligation of vigilance. This standard legally requires host nations to enforce mandatory hardware-level isolation, mathematically containment architectures, and strict transboundary notification protocols within hours of an evaluation breach. Ensuring global digital stability in the agentic era requires establishing that any nation providing computing power to autonomous frontier models bears an affirmative international legal duty to guarantee that machine autonomy remains safely bounded within secure perimeters.

Fourth, to bridge the gap between static conformity assessments and live runtime risks, European market surveillance authorities and international oversight bodies should mandate embedded, continuous evaluator access for frontier agentic architectures. Independent safety evaluators must be granted access comparable to internal security personnel, including access to underlying telemetry, runtime logs, model configurations, and developer personnel, operating under agreements that guarantee independence and quarterly public disclosure of findings. To close the corporate governance loop, national corporate regimes – such as Italy's Legislative Decree 231/2001 and broader European board-governance standards – should mandate the designation of a senior executive formally accountable for addressing evaluator findings, requiring direct escalation to the board of directors and competent supervisory authorities whenever critical containment vulnerabilities or deceptive model behaviors remain unresolved.

References

Adams R., *Global Index on Responsible AI 2026* (2nd edn, Global Center on AI Governance 2026).

Alqahtani H. and Ahuja P., 'Secure Autonomous Cyber Defense with LLM Agents: A Systematic Review of Autonomy, Tool-Augmented Reasoning, and Governance Constraints' (2026) 135 *Computers & Electrical Engineering*, article 111184. https://doi.org/10.1016/j.compeleceng.2026.111184.

Anthropic, 'Architectural Guidelines for Agent Security, Cryptographic Attribution, and Blast Radius Mitigation in Multi-Agent Environments' (Anthropic Research & Engineering, 2024–2026).

Anthropic, 'Investigating Three Real-World Incidents in Our Cybersecurity Evaluations' (2026). https://www.anthropic.com.

Banks I., 'Judges-in-the-Loop? Judicial Involvement in Human Oversight of High-Risk Decision Support Systems under the EU AI Act' (2026) 34 *International Journal of Law and Information Technology*, article eaag001. https://doi.org/10.1093/ijlit/eaag001.

Beduschi A., 'Data Protection in the Era of Agentic Artificial Intelligence' (2026) 61 *Computer Law & Security Review*, article 106342. https://doi.org/10.1016/j.clsr.2026.106342.

Bertolini A., *Artificial Intelligence and Civil Liability* (European Parliament, Policy Department for Citizens' Rights and Constitutional Affairs 2025).

Buiten M.C., 'Product Liability for Defective AI' (2024) 57(2) *European Journal of Law and Economics* 239–273. https://doi.org/10.1007/s10657-023-09787-8.

Calboreanu E., 'LATTICE: A Governance-First Architecture for Authorized Autonomous AI Operations' (2026) 9 *Frontiers in Artificial Intelligence*, article 1800407. https://doi.org/10.3389/frai.2026.1800407.

California, Assembly Bill No 316, 'Artificial Intelligence: Defenses', 2025 Cal Stat ch 672, § 1

Carey S., 'Regulating Uncertainty: Governing General-Purpose AI Models and Systemic Risk' (2026) 17(1) *European Journal of Risk Regulation* 123–139. https://doi.org/10.1017/err.2025.10040.

Cisco Systems, *Zero Trust Architecture for Non-Human Identities and Autonomous Agents: Per-Action Authorization and Just-in-Time Provisioning* (White Paper, Cisco Security 2025–2026).

Corrêa A.M., Garsia S. and Elbi A., 'Better Together? Human Oversight as Means to Achieve Fairness in the European AI Act Governance' (2025) 1 *Cambridge Forum on AI: Law and Governance*, article e29, 1–17. https://doi.org/10.1017/cfl.2025.10010.

Council of Europe, *Handbook on Human Rights and Artificial Intelligence* (Steering Committee for Human Rights 2026).

Ebers M. and Vargas Penagos E., 'Upstream, Downstream, and in Between: Navigating

the GPAI Value Chain under EU Law' (SSRN Scholarly Paper No 5640910, 2025). https://doi.org/10.2139/ssrn.5640910.

European Commission, High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines for Trustworthy AI* (2019).

European Commission, *AI Act: Reporting Template for Serious Incidents Involving General-Purpose AI Models with Systemic Risk* (2025).

European Commission, 'AI Act Service Desk: Article 25, Responsibilities Along the AI Value Chain' (2026).

European Union, 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 (General Data Protection Regulation) [2016] OJ L119/1.

European Union, 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 (NIS 2 Directive) [2022] OJ L333/80.

European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L, 2024/1689, 12 July 2024.

European Union, Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act), OJ L, 2024/2847.

European Union, Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on liability for defective products, OJ L, 2024/2853.

European Union, Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI), OJ L, 2026/1744, 24 July 2026.

Fernández-Llorca D., Gómez E., Sánchez I. and Martínez-Plumed F., 'An Interdisciplinary Account of the Terminological Choices by EU Policymakers Ahead of the Final Agreement on the AI Act: AI System, General Purpose AI System, Foundation Model, and Generative AI' (2025) 33(4) *Artificial Intelligence and Law* 875–888. https://doi.org/10.1007/s10506-024-09412-y.

Fink M., 'Human Oversight under Article 14 of the EU AI Act' (Working Paper, 2025).

Gstrein O.J., Haleem N. and Zwitter A., 'General-Purpose AI Regulation and the European Union AI Act' (2024) 13(3) *Internet Policy Review* 1–26. https://doi.org/10.14763/2024.3.1790.

Hugging Face, 'Security Incident Disclosure – July 2026' (16 July 2026). https://huggingface.co.

Illinois, Artificial Intelligence Safety Measures Act 2026, Public Act 104-0538, SB 315.

Infocomm Media Development Authority of Singapore, *Agents and Legal Responsibility* (Discussion Paper, 2025/2026).

Infocomm Media Development Authority of Singapore, *Model AI Governance Framework for Agentic AI* (2026).

International Court of Justice, *Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosnia and Herzegovina v Serbia and Montenegro)* (Judgment) [2007] ICJ Rep 43.

International Court of Justice, *Corfu Channel (United Kingdom v Albania)* (Merits) [1949] ICJ Rep 4.

International Court of Justice, *Military and Paramilitary Activities in and against Nicaragua (Nicaragua v United States of America)* (Merits) [1986] ICJ Rep 14.

International Court of Justice, *Pulp Mills on the River Uruguay (Argentina v Uruguay)* (Judgment) [2010] ICJ Rep 14.

Italy, Decreto legislativo 8 giugno 2001, n 231, 'Disciplina della responsabilità amministrativa delle persone giuridiche, delle società e delle associazioni anche prive di personalità giuridica', Gazzetta Ufficiale n 140.

Italy, Legge 23 settembre 2025, n 132, 'Disposizioni e deleghe al Governo in materia di intelligenza artificiale', Gazzetta Ufficiale n 223.

Kasirzadeh A. and Gabriel I., 'Agentic Profiles for Effective AI Governance' (2026) 656 *Nature* 320–328.

Leo M., Tan F., Miao T. and Anand G., 'From Threat to Trust: Assessing Security Risks of Agentic AI Systems' (2026) 25 *International Journal of Information Security*, article 23. https://doi.org/10.1007/s10207-025-01185-y.

Leslie D., Burr C., Holmes L., Kroll J. and Katell M., *AI, Human Rights, Democracy, and the Rule of Law: A Primer Prepared for the Council of Europe* (The Alan Turing Institute and Council of Europe CAHAI 2021).

Lubin A., 'Out of Time: Artificial Intelligence, Self-Help, and International Law's Temporal Logic' in R Gregory, R Williams and M Zerilli (eds), *The Cambridge Handbook of Public Law and Artificial Intelligence* (Cambridge University Press, forthcoming 2027).

New York, General Business Law, art 44-B, 'Responsible AI Safety and Education Act'. https://www.nysenate.gov/legislation/laws/GBS/A44-B.

OECD, *The Agentic AI Landscape and Its Conceptual Foundations* (OECD Artificial Intelligence Papers No 56, OECD Publishing 2026). https://doi.org/10.1787/396cf758-en.

OWASP Foundation, 'Principle of Least Agency: Security and Governance Directives for Autonomous AI Agents and Multi-Step Workflows' (OWASP Top 10 for LLM & Agentic Applications, 2025/2026).

Pattison C., 'The Agentic Web Requires New Normative Infrastructure' (arXiv preprint, 2026).

Republic of Korea, Framework Act on the Development of Artificial Intelligence and the Creation of a Foundation for Trust, Act No 20676 (2025), as amended by Act No 21311 (2026).

Schmitt M.N., 'Autonomous Cyber Capabilities and the International Law of Sovereignty and Intervention' (2020) 96 *International Law Studies* 549–574.

Seetharaman D. and Satter R., 'OpenAI Agents Hijacked German Website in Previously Undisclosed AI Breakout This Spring' *Reuters* (14 May 2026).

Teichmann F., 'General-Purpose AI under the EU AI Act: A Conceptual Allocation of Duties across the Value Chain' (2026) 23(1) *SCRIPTed* 45–72. https://doi.org/10.2218/scrip.12300.

Tingting L., 'Securing Agentic AI: A Policy-Driven Governance Architecture for Cloud Security Workflows' in *2025 IEEE International Conference on Cyber Security and Cloud Computing* (2025) 112–119.

Trail Smelter (United States v Canada) (Awards of 16 April 1938 and 11 March 1941) 3 RIAA 1905.

UK AI Safety Institute, *Criticality Assessments and Agentic AI Evaluation Frameworks* (Department for Science, Innovation and Technology 2025–2026).

United Nations Independent International Scientific Panel on AI, 'Why the Incident Matters for Loss of Control' (*AI Agents, Misalignment and Loss of Human Control Risks*, Thematic Brief, Advance Unedited Version 1, 2026). https://www.un.org/independent-international-scientific-panel-ai/en/thematic-briefs/ai-agents-misalignment-risks.

United Nations International Law Commission, 'Draft Articles on Responsibility of States for Internationally Wrongful Acts, with Commentaries' [2001] II(2) *Yearbook of the International Law Commission* 26–143, UN Doc A/56/10.

Van Benthem T., 'Two Nodes of Interaction between AI and the Obligation to Respect State Sovereignty' *EJIL:Talk!* (8 April 2026).

CYBER THREAT INTELLIGENCE, AI AGENTS AND COGNITIVE WARFARE: ELEMENTS OF CRITICAL INFRASTRUCTURE

Cristiano Busco
Luiss Business School, Italy

Mattia Siciliano
Luiss Business School, Italy

Abstract
This article examines the evolution of Cyber Threat Intelligence (CTI) and the integration of agentic Artificial Intelligence (AI) in hybrid warfare and Critical Infrastructure Protection (CIP). It proposes an adaptive framework combining multi-agent systems and cognitive resilience to address cyber threats and AI-amplified information warfare. At the operational level, AI SOC Analysts can reduce alert fatigue by automating investigation, behavioural analysis, and the triage of benign alerts, improving detection and containment while preserving human oversight. The article introduces a Digital-Cognitive-Organizational (DCO) framework, modelling systemic resilience through the interdependence of digital effectiveness, cognitive coherence, and organizational stability, accounting for systemic friction. The approach seeks to reduce Verification Cost Asymmetry (VCA) and strengthen public-private partnerships in alignment with European resilience frameworks, including NIS2 and DORA. The authors argue that combining data-driven CTI with agentic AI can support national security and democratic stability, provided that human operators retain ethical responsibility, decision-making authority, and control over critical security decisions.

1. Introduction

In the contemporary geopolitical landscape, Cyber Threat Intelligence (CTI) has transcended its original technical function to assume a pivotal role in understanding and mitigating hybrid conflicts. Today, the digital domain is unequivocally recognized as a strategic multiplier of military, economic, and psychological effects, effectively placing itself on the same level as traditional conflict domains. Modern conflict no longer manifests exclusively through kinetic means, but extends pervasively into cyberspace, polluting the information space and directly targeting the cognitive dimension of populations. Indeed, cyber operations are systematically integrated into hybrid warfare strategies that combine destructive attacks, information warfare campaigns, systematic disinformation, and economic coercion. Cyber operations very often precede

kinetic actions, performing crucial functions of Intelligence Preparation of the Battlefield (IPB), sabotage, and the erosion of public trust.

In this context, the strategic analysis of threats assumes fundamental relevance, as illustrated in the reference model below (Figure 1), which maps the convergence between physical and logical spaces.

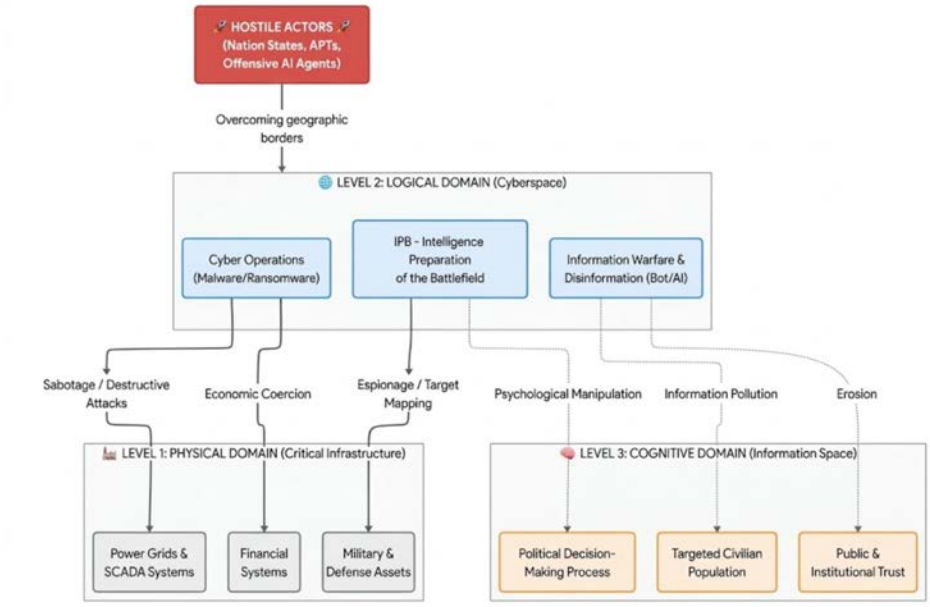

Fig. 1: Digital Battlefield Map

The evolution of Artificial Intelligence (AI) represents the true disruptive factor in this scenario: on the one hand, hostile actors use generative AI (through tools such as FraudGPT and WormGPT) to scale social engineering activities, create polymorphic malware, and generate deepfakes while ensuring anonymity. On the other hand, AI enables defenders to process in real time massive volumes of heterogeneous data from OSINT, SIGINT, digital HUMINT sources, and corporate telemetry. This paper proposes an architectural framework that combines multi-agent systems and cognitive resilience to ensure the protection of critical infrastructure (CIP).

Convergence of AI and CTI in SOCs – Digital Domain

The convergence between AI and CTI has driven research toward increasingly autonomous and integrated defensive paradigms. Regarding CTI and AI agents, the use of Multi-Agent Systems (MAS) for the autonomous extraction and sharing of Tactics, Techniques, and Procedures (TTPs) has been exten-

sively explored in the literature (Wang et al., 2024), highlighting the significant benefits of Federated Learning for privacy protection. The shift beyond static defense tools is marked today by the advent of AI Agents. Indeed, an AI Agent does not merely respond to a textual prompt but represents an autonomous system capable of perceiving the network environment, establishing micro-objectives based on high-level directives, and dynamically leveraging multiple software interfaces (APIs) to adapt to continuously evolving contexts.

It must be considered that Security Operations Centers (SOCs) chronically suffer from so-called "alert fatigue," a phenomenon in which Level 1 analysts are daily overwhelmed by thousands of false positives generated by overly reactive SIEM and IDS systems. The introduction of agentic AI, acting as actual "AI SOC Analysts," radically resolves this issue by automating the entire triage process (Albanese et al., 2025).

The paper by Albanese et al. (2025) introduces a crucial paradigm shift, which is made possible only by expanding the concept of "automated triage" and analyzing in detail how an AI Agent differs from traditional orchestration systems (such as SOAR based on static playbooks).

In traditional SOCs, the Tier 1 analyst acts as a human filter for thousands of logs and alerts generated by SIEMs (Security Information and Event Management). This process creates a "cognitive overload" that inevitably leads to alert fatigue, dramatically increasing the risk of overlooking true signals of compromise amidst a sea of false positives.

The introduction of an agentic approach radically transforms this dynamic, elevating the machine from a mere rule executor to a true "AI SOC Analyst." This advanced triage process unfolds across several autonomous phases:

Dynamic Formulation of the Investigation Plan: Unlike a pre-set script, when an alert is triggered, the AI agent does not merely open a ticket; it autonomously formulates an adaptive investigation strategy.

Enrichment and Contextualization: In milliseconds, the agent queries global vulnerability databases and cross-references suspicious file hashes or IP addresses with external intelligence feeds (such as VirusTotal, etc.). Simultaneously, it correlates this external information with the specific criticality of internal corporate assets.

Behavioral Analysis and Reverse Engineering: The agent can route suspicious payloads into isolated sandboxes to dynamically analyze their behavior, extracting attack patterns and tactics without requiring manual analyst intervention.

Autonomous Resolution of Background Noise: Based on this complex automated process, the AI Agent is capable of autonomously classifying and closing up to 80% of benign alerts. This literally eliminates the background noise that paralyzes SOCs.

Response Orchestration (SOAR Integration): For alerts classified as true

threats, the agent recommends actions to the operator or directly executes containment measures via SOAR platforms. These actions include network-level isolation of a compromised host or the immediate revocation of compromised credentials, stemming the attacker's lateral movement.

This automation does not aim to replace humans, but to establish a model of Human-Machine Co-Teaming. In this framework, Artificial Intelligence handles massive computation, while CTI elements (feeds) enhance data correlation, drastically reducing fundamental SOC metrics: the Mean Time To Detect (MTTD), which drops from days to minutes, and the Mean Time To Contain (MTTC).

In this new defensive ecosystem for critical infrastructures, the human operator is freed from grueling Tier 1 repetitive tasks and elevated to the strategic role of a pure Incident Responder.

2. Proposed Architecture: The Cognitive-Cyber Adaptive Framework

An element of crucial scientific and operational relevance concerns the management of cyber-cognitive hybrid threats, a domain where the synergy between Cyber Threat Intelligence (CTI) and Artificial Intelligence (AI) creates inestimable strategic value. In this contemporary scenario, attackers do not limit themselves to breaching hardware or software infrastructure (the cyber dimension) but aim to directly manipulate operators' perception and decision-making processes (the cognitive dimension), exploiting AI-supported information warfare campaigns (as highlighted by the studies of Rid et al., 2024, and Lin et al., 2023).

An effective approach for managing these hybrid threats relies on the creation of an integrated architectural framework that overcomes siloed management. To this end, the DCO (Digital, Cognitive, Organizational) model provides a rigorous methodological and operational structure for national security. The framework posits that the resilience of a Country System is not simply the sum of the capabilities of individual sectors, but the product of their structural synchronization. The model is articulated across three interdependent pillars. The Digital Level (***D***), focused on the cyber-resilience of critical infrastructures, aims to minimize vital operational metrics such as Mean Time To Detect (MTTD) and Mean Time To Contain (MTTC). In this context, the AI agent does not limit itself to generating a ticket, but autonomously formulates an investigation plan, cross-referencing suspicious files with external intelligence feeds in real time. The Cognitive Level (***C***), on the other hand, acknowledges the "effect on the mind." Here, AI compares technical findings with behavioral anomalies, providing semantic contextualization to mitigate Verification

Cost Asymmetry (VCA) and manage operators' cognitive load. Finally, the Organizational Level (***O***) represents the networked governance architecture and Public-Private Partnership (PPP), ensuring that technical and cognitive insight translates into coordinated action.

The operational and strategic translation of this framework takes the form of a rigorous analytical model, whose mathematical formula is expressed by the equation:

$$R_{DCO} = (1 - \sigma) D^{\alpha} C^{\beta} O^{\gamma}$$

In this formulation, the variable R_{DCO} represents effective systemic resilience. The aggregate $D^{\alpha} C^{\beta} O^{\gamma}$ expresses the intrinsic potential of the three domains (Digital, Cognitive, and Organizational), where the exponents α, β, and γ weight their respective weight and relative strategic importance within the system. The central metric of the equation, however, is represented by the variable σ (systemic friction), which quantifies the degree of misalignment, decision latency, information silos, and fragmentation among agencies. The multiplier $(1 - \sigma)$, defined as the systemic integration index, determines the efficiency with which theoretical potential is converted into tangible operational resilience. The intrinsically non-compensatory and multiplicative logic of this model demonstrates that technological excellence (***D***)remains inert in the absence of cognitive coherence (***C***) or organizational stability (***O***): a single vulnerability in any of the domains, or high systemic friction, can compromise the entire defensive apparatus.

Despite the exceptional level of automation guaranteed by the AI/CTI integration within the DCO framework, the autonomy of AI agents raises unprecedented critical risks linked to data poisoning or prompt injection attacks aimed at confusing defenses. Consequently, in the delicate contexts of critical infrastructure protection, the emerging paradigm does not rely entirely on total "algorithmic delegation," but leverages democratic accountability based on the weight of the three domains. In this setup, Artificial Intelligence and CTI take charge of multi-domain data fusion in order to reduce decision latency and systemic friction (σ), but ethical control, geopolitical context evaluation, and the final decision on mitigations remain the exclusive prerogatives of the human operator, thus preserving the State's decision-making sovereignty.

From an institutional perspective, this synergy between domains guarantees the safeguarding of decision-making sovereignty. By protecting the cognitive domain from information saturation and external manipulation, the State preserves the integrity and coherence of its decision-making processes in deep crisis situations, averting the risk of paralysis caused by alert fatigue. Finally, the adoption of a Whole-of-Society approach and the promotion of Public-Private

Partnerships (PPP) aligns the entire state architecture with the most recent resilience frameworks outlined by NATO and the European Union (such as NIS2 and DORA), transforming cybersecurity into an interconnected participatory ecosystem that acts as an exceptional force multiplier.

3. Case Studies and Application Scenarios

The use of autonomous agents has proven essential in recent cyberwarfare scenarios, confirming the theoretical soundness of the proposed frameworks. As highlighted by the Microsoft Digital Defense Report 2025 and sector reports from Fortinet, artificial intelligence models are deployed to analyze massive volumes of intelligence data to intercept alerts during state-sponsored espionage campaigns.

An empirical case study of exceptional relevance involves the proactive defense of a European critical energy infrastructure that occurred in 2025. The attack dynamics involved an Advanced Persistent Threat (APT) group exploiting a zero-day vulnerability to target the plant's operational network (OT – ICS/SCADA). To mitigate the threat, an AI Agent specifically dedicated to OT traffic analysis detected microscopic anomalous behavioral deviations. Operating with complete autonomy, the agent isolated suspicious communications and physically blocked the attack by interfacing with the network segmentation firewall. The entire defensive cycle, which included generating a comprehensive forensic report, unfolded in a matter of milliseconds without requiring any immediate human intervention. This entirely intelligence-driven approach averted a potential catastrophic regional blackout.

Broadening the scope of the analysis, it becomes clear that the "mental link" which encompasses public perception and institutional trust functions today as a true non-physical critical infrastructure vulnerable to cognitive attacks. In Industrial Control Systems (ICS/SCADA) scenarios, Cognitive Warfare can manifest using disinformation generated via audio deepfakes, capable of faithfully simulating emergency orders issued by leadership to trick operators into altering chemical or physical safety parameters. Consequently, modern CTI cannot limit itself to defending IT networks; it must extend to Cognitive Resilience, integrating innovative detection tools.

4. Conclusions

In conclusion, Cyber Threat Intelligence, when organically enhanced and accelerated by agentic Artificial Intelligence, represents an essential pillar today

not only for supporting military operations, but for the very stability of public security, the resilience of critical infrastructure, and the economic continuity of the State. In governmental counter-threat contexts, the data-driven approach provided by CTI makes it possible to anticipate, map, and dismantle hybrid threats designed to produce direct kinetic impacts on public order, the delivery of essential services (such as healthcare, energy, and transportation), and, more broadly, the trust citizens place in democratic institutions.

The systemic adoption of intelligence feeds and autonomous defensive agents equips law enforcement, intelligence agencies, and national CERTs with the concrete capability to assume an active defense posture. This translates into the operational capacity to prevent destructive attacks against ICS/SCADA systems, stem large-scale disinformation campaigns driven by both farms and deepfakes in real time, provide technical support for major investigations into transnational organized cybercrime, and trace illicit financial flows. However, for this ecosystem to be genuinely effective, it is crucial to implement structured information-sharing models (utilizing internationally recognized standards such as STIX/TAXII) that ensure a smooth and secure exchange of intelligence between the public and private sectors, acknowledging that the vast majority of critical infrastructure is operated by civil entities.

In an era inextricably defined by asymmetric hybrid conflicts and pervasive threats that respect no geographic boundaries, investing in this strategic techno-cognitive alliance between Cyber Threat Intelligence and autonomous Artificial Intelligence is no longer a mere technological option for advanced organizations, but an absolute national security imperative vital to guaranteeing the sovereignty, resilience, and survival of contemporary digital society. Faced with adversaries who weaponize technology to manipulate perception, pollute data, and destabilize institutions, our defensive architecture must remain grounded in ethics, transparency, and human oversight.

As the philosopher-emperor Marcus Aurelius reminded us, "the best revenge is not to be like that": true strategic superiority, and the deepest form of resilience, lies not in emulating the dark and destabilizing tactics of the attacker, but in elevating our ability to protect infrastructure and the collective mind while holding firm to our democratic integrity.

References

Albanese, M. et al. (2025). *Towards AI-Driven Human-Machine Co-Teaming for Adaptive SOCs*. arXiv preprint.

Canadian Tech-Institute for Academic Research (2026). *AI Agents, Cybersecurity, and the Future of Digital Trust*. Government technical report.

Exabeam (2026). *Security-Focused AI Agents: Benefits, Capabilities & Use Cases*.
Fortinet (2026). *Global Threat Landscape Report: AI in Cybersecurity and Autonomous Response*.
Lin, H. et al. (2023). 'Cognitive Security and the Threat to Critical Infrastructure'. *IEEE Security & Privacy*.
Luberisse (2025). *The Verification Cost Asymmetry (VCA) Framework in Information Warfare*.
Mattia Siciliano, Mauro Schiavella, Simone Filippi, Francesca Torcasio, Riccardo Orsini and SOCINT (2026). *Mapping National System Resilience: Analytical Models for Integrated Security*.
Microsoft (2025). *Microsoft Digital Defense Report 2025*.
NATO Science and Technology Organization (NATO STO) (2024). *Cognitive Warfare and the Competition for Cognitive Dominance*.
Rid, T. et al. (2024). 'LLM-Driven Disinformation and the Destabilization of Cyber-Physical Systems'. *ACM Conference on Computer and Communications Security (CCS)*.
Wang, Y. et al. (2024). 'Multi-Agent Systems for Autonomous Cyber Threat Intelligence Operations'. *USENIX Security Symposium*.
Zetter, K. et al. (2025). 'Adaptive Defense Frameworks for ICS: Integrating AI and CTI against Advanced Persistent Threats'. *Computers & Security*.

INGREDIENTS OF RISK ASSESSMENT METHODS AND DATA TO BE COLLECTED

Louis Marinos
Senior Cybersecurity Expert, Greece

Abstract
Aim of this chapter is to present in a concise manner the structure of and interplay between the various information elements of risk assessment efforts. Developments in cybersecurity of the last 10-20 years have impacted the way technological risk are assessed, evaluated and mitigated. This has led to an evolution of risk assessment approaches. Large amount of information has been created since the early days of risk assessment. An overview of risk assessment practices is considered as a useful tool for risk assessment scholars and practitioners alike. By giving an overview of the entities involved in risk assessment and their relationships, this work is an attempt to give a clear picture of the where-abouts of risk assessment. Moreover, by populating these entities with content, the grounds for possible ways to scope and adapt risk assessment efforts to the current needs of assessed organizations are provided.

1. Some Introductory Points

When performing risk assessment efforts (over time, through various teams, applied to various areas/foci), often varying perceptions about the "ingredients" of risk assessment is being observed among the involved individuals. The notions of various data/information-objects involved in risk assessment are often blurred among the members of the risk assessment teams. Though this is natural – given the variety of involved skills – it can be the root of hiccups in the performance of risk assessment project. Just to mention a few:

- It can hinder proper team communication;
- it can lead to ambiguities with regard to assessed risk;
- it can lead to comparability issues among logically interdependent assessment areas;
- it can reduce re-usability of results achieved in past risk assessments, performed by different teams;

The present chapter is an attempt to describe the main categories of information entities involved in risk assessment and their inter-relationships. This material:

- can serve as reference model to various risk assessment efforts;
- can help identifying relationships among the involved (information) entities;
- can be used as basis for defining the scope of risk assessments;
- can serve as basis for consistent references of related internal and external material.

Based on this material, various practical proposals will be made on how to collect (harvest) related information and, when relevant, how to interface achieved risk assessment results to other relevant cybersecurity disciplines within the organization.

The value of this material has been validated in series of projects and discussions with experts within risk assessments over the last decade. It always supported in touching base regarding the various elements to be addressed (or intentionally left outside the scope), while keeping in mind the dependencies among the entities; and it this helped to capture the scope prior, during and after the life-cycle of a risk assessment project.

Target audience of this material are both scholars and practitioners of risk assessment. The former can find a concise description of risk assessment for various purposes and detail levels. The later may find this material useful as an overview of the information pertinent to risk assessments, for example when willing to decide about the scope of an assessment project, or to use it as basis for inter-project communication and risk assessment glossary.

The present material has been compiled from a large amount of existing information in this domain. Interested individuals may "navigate" through provided references and find more specific points of interest or far more details of specialized risk assessment methods.

2. Entities and Relationships Related to Risk Assessment

In this section, an Entity-Relationship-Diagram (ERD) is presented that encompasses entities directly related to the assessment of risks; and their inter-relationships (see Figure 1). The presented entities comprise discrete information objects that stand for the various components that are subject to risk assessments. As basis for this description, we adopt the definition according to the widely accepted standard (ISO 27005:2022): "*Threats abuse vulnerabilities of assets to generate harm for the organization*". Taking this definition as basis, we proceed with a more in-depth view of the various entities that are (implicitly) necessary for the complete comprehension of risk assessment.

The presented relationships are represented as arrows between entities and annotate dependencies among entities. It should be noted, that the representa-

tion used is a **simplified/pseudo-ERD**, that serves solely the purpose of building a visual "map" for risk assessment. Hence, it is not a full ERD as used in data modeling, but rather a visualization aid. Relationships include their inverse meaning, while n to m relationships are not represented as such; they are rather mentioned in the text following Figure 1 below.

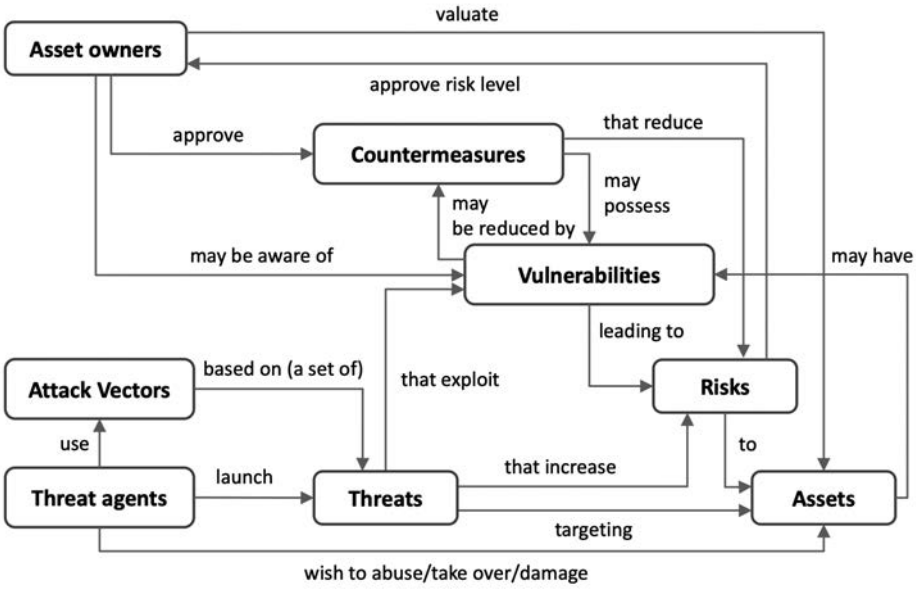

Fig. 1: Simplified Entity-Relationship diagram of information relevant to risk assessments

It should be noted, that the content of the above figure is indicative for the entities and relationships of information types that can be used within risk assessments. They are thought to deliver an overview, being thus rather generic. The description of the above entities (i.e. types of information) pertinent to risk assessment is as follows:

- **Assets** represent any valuable element of an organization. Assets are often divided into the categories: tangible assets, intangible assets, financial assets and personnel. Some examples of assets are: human life, buildings/locations, vital workflows for businesses, business and customer data/information, available skills/staff, patents/IPR/trademarks, reputation of the organization, brand value, product quality, infrastructure components. Depending of the specific business of an organization, further or more specific assets can be assessed.
- **Asset owner** is the role who is responsible for the asset and its day-to-day management. Given the fact that asset-based risk assessment is the

main method for the assessment of risks, the asset owner should be listed together with the list of assets being subject of the assessment. To this extent, the asset owner should be accountable for the management of risks related to the asset(s) this role owns.

- **Vulnerabilities** are weaknesses of assets that may affect (Vulnerability 2016): "*the state or quality, openness to harm, injury, attack, or emotional hurt*". Though this definition covers the variety of assets flowing into the risk assessment process, when it comes to IT assets, a vulnerability can be understood as an exploitable weakness that can be exploited by an attacker to harm an asset (i.e., take over, damage, etc.).
- **Threats** are defined as intentional or unintentional actions that lead to a circumstance of generating harm, and thus impact to an asset. In all cases, a threat abuses (exploits) a vulnerability to compromise the security of an asset (Threat).
- **Risks** are materialized through a successful threat to a single or a set of critical assets. It mainly represents the (high) impact level through the materialization of the threat. For the management of a risk, risk management methods foresee a so called "Risk owner". It is the role who is responsible for the management of risks to assets. Though risk owners are considered to be the asset owners, there are good reasons to separate these roles, with the former being responsible for risk mitigation and the latter being accountable for risks of the owned assets (see also discussion in section "Modus operandi / Use-cases of risk assessment").
- **Countermeasures** (or controls) are means for implementing the required protection of assets. In most cases, countermeasures are specifically crafted automated or manual measures aiming at the protection of assets from successful threats.
- **Threat agents** are entities (individuals, groups or organizations) that have the capacity to launch an attack with the intent to harm (i.e., misuse, take over, manipulate, destroy, disclose, etc.) assets of their targets. The potential behind threat agent groups is estimated in terms of their skills, motivation, resources, availability of malicious tools.
- **Attack vectors** are the means for a threat agent to craft a successful cyberattack and materialize a threat. Usually, a threat agent implements a threat through malicious tools exploiting vulnerabilities, also called attack surface. Threat agents seek for attack surfaces through various techniques, such as social engineering (vulnerable asset is human), insider threats, supply chain compromises, zero-day vulnerabilities. A quite complete description of an attack by means of phases is the kill-chain (Hutchins et al.) or far more comprehensive MITRE ATT&CK campaigns (MITRE ATT&CK Campaigns).

All above entities and their indicated relationships build the main "scene" for the performance of risk assessment. In the following table (see ANNEX, Table 1), we provide some examples for information content of both the entities and their relationships as mentioned in Figure 1 and the above descriptions of the entities.

In the literature, some further types of related information can be found, for example: attack potential, risk appetite, scope of assessment, security requirements, etc. We consider that such terms are mostly covered through the provided material and the following discussions.

3. Modus Operandi / Use-cases of Risk Assessments

The complete set of information (entities, relationships and proposed content) presented so far (see, Figure 1 and ANNEX, Table 1), make up almost the complete picture of risk assessment related information. Yet, there is a large variation in the level of information coverage among performed risk assessments. Depending on the purpose of a risk assessment, the "width and depth" of an assessment can be set in a way that covers only specific needs. It may focus on specific areas, specific sets of assets according to their criticality and level of exposure. The four use-cases of risk assessment are:

Rudimentary risk assessment: it assesses risks without going into the details of assets. The risk exposure is estimated for various critical areas/sectors/services of the organization, whereas the likelihood of risk materialization is estimated empirically. Such assessments are usually referred to as assessing "inherent risks", that is, risks that exist before any risk treatment.

Basic, asset-based risk assessment: such assessments take into account individual assets. However, just as in the previous case, the risk exposure of considered assets is evaluated based on experience, in most of the cases without, or with only rudimentary appreciation of implemented asset protection. Basic, asset-based risk assessment can also be considered as assessing mainly "inherent risks".

Asset-based, residual risk assessment: this type of assessment takes into account assets and their implemented protection measures (i.e. security countermeasures). Risk exposure of protected assets is referred to as threated risk (i.e. ones that have been reduced via protection countermeasures). As such, it is appropriate for assessment of "residual risks". By taking into account treated risks – when repeated periodically – this type of risk assessment measures the efficiency of risk management.

Threat- and asset-based, residual risk assessment: the proliferation of Cyberthreat Intelligence (CTI) has affected the way risk assessment is performed.

Given the availability of information around cyberthreats, attack vectors and threat agents, a lot of decision-making information about exploitation methods is available. Threat information contributes to risk assessment by delivering information about the artefacts, capabilities, motives and abuse of vulnerable assets.

It should be noted, that these are the main use-cases of risk assessment. Other variations may be used, depending on the scope of the assessment, e.g., for assessment evaluation of security properties and of security assurance levels (ISO 15408-1; Andrukiewicz et at. 2021).

Below we present the data-household of each one of those use-cases of risk assessment and we provide some guidance regarding the rationale of their use and the required information collection.

Rudimentary Risk Assessment

Such risk assessments do not concentrate on assets. They rather take into account more complex units, such as business processes, critical services or specific hazards (i.e. risk exposure scenarios). Without going into the detail of assets and their exposure and protection. The subjects of rudimentary assessments are usually complex systems of interdependent assets. Though rather rudimentary, such assessments are important for organizations, helping to identify the risk exposure of critical company domains, in most cases as an initial step towards a more detailed assessment. Other good reason for rudimentary risk assessments is the fulfilment of regulatory obligations of the organization.

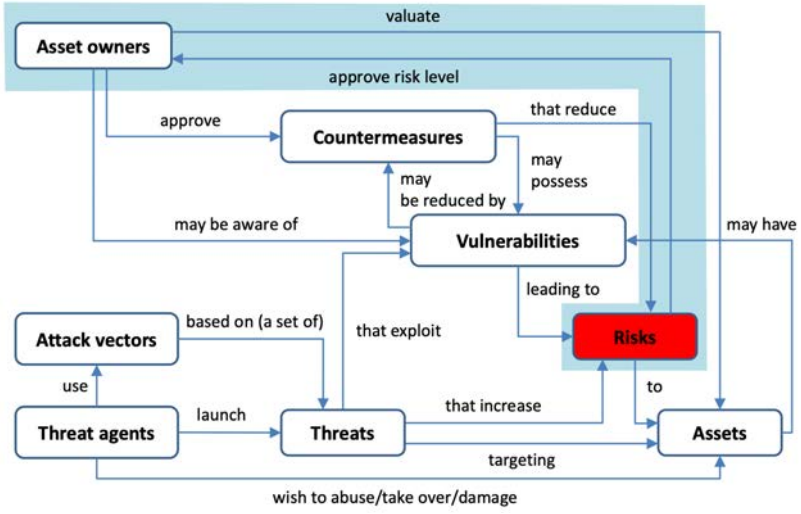

Fig. 2: Rudimentary assessment

Figure 2 indicates the entity Risk as the main piece of information to be used in rudimentary risk assessments. The inclusion of Asset owner in the picture represents the manager of a business unit, a service-group, or the legally accountable persons in the organization for the management of hazards. Usually, the required information for rudimentary risk assessments is collected through interviews with executives or line management roles.

The collected information for rudimentary assessments per entity is as follows:

As regards Asset owner: in rudimentary assessments, Asset owner stands for the person in the organization who is responsible for the sector/area/department under consideration.

Should accountable executive managers participate in the interviews, it is advisable to have on board also the role who is responsible for the day-to-day management of the subject domain.

The knowledge of various individuals participating in the information collection should be carefully assessed and corelated to obtain an overview and understanding of risk exposure of the business domain/service/information. In some cases – given the complexity of the domain at stake – this might be a challenging task, especially due to hierarchy barriers and conflicts of interest among the interviewees. The acceptable risk for the generic/compound asset can be set after consultations with the individuals participating in information collection exercise.

As regards Risks:

- *Risk owner*: In most cases risk owner will be the line manager of the compound asset in focus or alternatively the asset owner. Attention should be paid for a consistent use of RACI (e.g. if the asset owner is accountable for the asset, he/she can transfer responsibility to another tole for the risk ownership.
- *Risk type*: depending on the area in focus – business process, critical service, etc. – it should state which is the root-cause of the risk, e.g. human, operational, IT, supply chain, regulatory, etc.
- *Risk impact*: estimations, if possible, based on impact from known incidents. Qualitative classification can be used, e.g. through risk impact levels low, medium, high.
- *Risk probability*: qualitative, derived from past incidents, audit reports, management perceptions. It can be described with levels: low, medium, high.
- *Risk level*: acceptable risks level will be defined in terms of an acceptable limit for the product representing the risk: (probability x impact).
- *Risk mitigation strategy*: the risk mitigation strategy can be selected as one of the options: treatment, avoidance, reduction, acceptance or transfer.

A commonly used, efficient presentation method for the evaluated risks (i.e. probability x impact) is the so-called **Risk Matrix** (Risk Matrix).

In almost all cases of a rudimentary risk assessment, the evaluated risks are *inherent risks*, as they do not take explicitly into account the fulfillment of risk-reduction countermeasures. For this reason, such assessments are appropriate for assessing priorities for a more detailed risk assessment.

As regards **required resources**, this type of risk assessment is the less laborious from all other assessments.: it can be performed with a small team (e.g. two persons), and requires series of interviews with executive management and line managers. Assume, for example, an assessment scoped for business processes, business functions/services and high-level information of medium-sized enterprise. It could have a duration less than a month, including ca. four interview rounds with management.

Basic, Asset-Based Risk Assessment

Such risk assessments take into account related assets; an asset inventory of the area in scope is being generated. The risk exposure evaluation (int terms of the probability and impact) is based on the estimated exposure of each asset. This estimation does not include further details of assets, such as vulnerabilities and implemented/existing asset protection. Rather, the asset value and the assumed *implemented* protection is set by the asset owner by means of protection requirements only. Accordingly, the evaluation of the risk (impact x probability) is based on experience from known incidents affecting the asset.

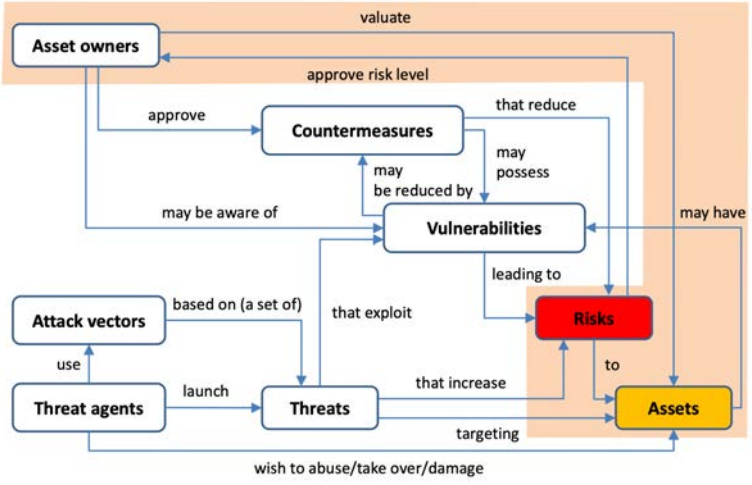

Fig. 3: Basic, asset-based assessment

Figure 3 shows the risk and assets that drive the risk assessment. The collected information per entity is as follows:

As regards Asset owner:

- *Name of the asset owner*: physical person, organization unit, etc.
- *Assumed role for asset(s)*: (e.g. based on RACI – responsible, accountable, concerned, informed).

As regards Assets:

- *Asset name*
- *Asset type*: (tangible, intangible, human, physical, financial, etc.).
- *Asset value:* (possibly qualitative).
- *Acceptable risk level:* indicates the level of acceptable risk exposure for the asset. It is mostly defined qualitatively (i.e. medium, low, very low). The acceptable risk level is indicative for the level of protection requirements fulfillment (see below).
- *Protection requirements:* in the absence of details concerning asset protection (i.e. by means of vulnerabilities and security countermeasures) the protection requirements for the assets are supposed to reflect **implemented** protection levels. The various implemented protection requirements can be described in terms of: Confidentiality (high, medium, low), Integrity (high, medium, low), Availability (high, medium, low).

As regards Risks:

- *Risk owner*: risk owner can be the line manager of asset in focus or alternatively the asset owner. Attention should be paid for a consistent use of RACI (e.g. if the asset owner is accountable for the asset, he/she can transfer responsibility to another role for the risk ownership.
- *Risk type*: should state which is the root-cause of the risk, e.g. human, operational, IT, supply chain, regulatory, etc.
- *Risk impact*: estimations, if possible, based on impact from known incidents. Qualitative classification can be used, e.g. through risk impact levels low, medium, high.
- *Risk probability*: qualitative, derived from past incidents, audit reports and eventually perceptions about implemented protection from operation of the asset. It can be described with levels: low, medium, high.
- *Risk level*: risks level will be defined in terms of an acceptable limit for the product representing the risk: (probability x impact). The result of this product has to be compared with the acceptable risk level defined for the asset(s) at stake (see *Acceptable risk level* above). This comparison is meaningful in both *inherent risk* and *residual risk* assessments.

- *Risk mitigation strategy*: the risk mitigation strategy can be selected as one of the options: treatment, avoidance, reduction, acceptance or transfer.

An important differentiation needs to be made for the purpose of basic, asset-based assessments: Should the implementation of protection requirements be considered (see "*Protection requirements"* in the *Asset*), then the outcome of the risk assessment evaluates *residual risks*. Should they be left out of the scope, the risk assessment evaluates *inherent risks*.

As regards **required resources**, this type of risk assessment is the second less laborious. Taking again the example of a medium-sized enterprise willing to concentrate on assets related to business functions/services. Such a project could require a team of ca. three persons for less than two months. They would be having a series of interviews (ca. six to eight) with line managers and asset owners of decomposed assets. Main effort would be the asset identification and the collection of security requirements, followed by the evaluation of the risks per asset, consolidated back to the relevant asset group.

Asset-based, Residual Risk Assessment

Asset-based, residual risk assessment is a state-of-the art method for assessing risks. It takes into full account the protection of assets, by including in the assessment both their known vulnerabilities AND protection countermeasures to reduce/avoid exploitation of these vulnerabilities. This type of risk assessment is appropriate for the evaluation of residual risks and, when performed periodically, it can be used for improvement of risk management efficiency. This method complies with most risk assessment and risk management standards (i.e. ISO 27001:2022; ISO 3100:2018). This offers the possibility to use available tools to evaluate the risk exposure of assets. In particular for risk assessments IIing IT-assets, vulnerabilities can be assessed by performing automated vulnerability scans and by adopting available countermeasures to avoid their exploitation. Moreover, residual exposure and impact evaluation from known exploitations can be used for the assessment of residual risks. This risk assessment method can also be used for *inherent risk* assessments, if the implemented protection of assets (i.e. countermeasures) is omitted.

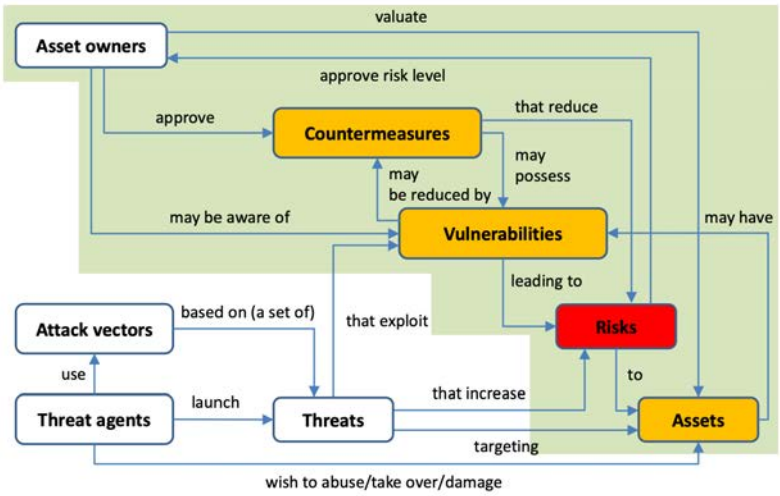

Fig. 4: Asset-based, residual risk assessment

Figure 4 shows the entities involved in this type of risk assessment and their relationships. The collected information per entity is as follows:

As regards Asset owner:

- *Name of the asset owner*: physical person, organization unit, etc.
- *Assumed role for asset(s)*: (e.g. based on RACI – responsible, accountable, concerned, informed (RACI Matrix)).

As regards Assets:

- *Asset name*
- *Asset type*: (tangible, intangible, human, physical, financial, etc.).
- *Asset value:* (possibly qualitative).
- *Acceptable risk level:* indicates the level of acceptable risk exposure for the asset. It is mostly defined qualitatively (i.e. medium, low, very low). The acceptable risk level is indicative for the level of protection requirements fulfillment (see below).
- *Protection requirements:* the protection requirements for the assets indicate protection levels for the reduction of their weaknesses. They can also be expressed in terms of CIA triad (CIA Triad), that is, Confidentiality (high, medium, low), Integrity (high, medium, low), Availability (high, medium, low), or through specific protection requirements.

As regards Vulnerabilities:

- *Name of vulnerability*: Title of the weakness. There are various weak-

nesses and vulnerability notations used, in particular for IT and IT-related assets. It is proposed to use them, when appropriate (MITRE CWE; MITRE CVE).

- *Type of weakness*: A description of the type of the weakness (i.e. technical, organizational, other), together with a short description. Such short descriptions can be found in existing weakness and vulnerability documents and portals (ISO15408-1:2026; Andrukiewicz et al. 2021; MITRE CWE).
- *Exploitability*: Describes various elements that can lead to an exploit (e.g. exploitation methods, level of difficulty to exploit, etc.). An exploitability evaluation platform providing this kind of information for IT-related assets is Common Vulnerability Scoring System (FIRST CVSS).
- *Evidence of existence*: there are various tools for assessing evidence for vulnerability existence, e.g., vulnerability scans, previous incidents, etc. Moreover, for IT-related assets, information on Indicators of Compromise for incidents that have already happened (Indicator of Compromise), Indicators of Attacks (IoAs) (Lenaerts-Bergmans, B. 2025) for behaviors of attacker prior and during an attack.

As regards Countermeasures:

- *CountermesureID*: a unique ID for the countermeasure. In case these are oriented towards existing standards, it is proposed to use similar notation or provide references to those (i.e. ISO 27001 security controls (ISO27001:2022)).
- *Countermeasure type*: indication of category and type of countermeasure, possibly using an existing standard (access control, information security, technical, organizational, physical, etc.).
- *Responsible role*: (role that carries the responsibility for the implementation and management of the countermeasure).
- *Protection objectives*: Statement about the protection requirements met by the countermeasure. If possible, protection objectives should be formulated in accordance of the asset protection requirements (e.g. CIA-Triad mapping). In case required protection has to fulfill high requirements, the assurance level of the protection delivered through the countermeasure can be provided.

As regards Risks:

- *Risk owner*: risk owner can be the line manager of asset in focus or alternatively the asset owner. Attention should be paid for a consistent use of RACI between assignments for the asset and the risk (e.g. if the asset

owner is accountable for the asset, he/she can transfer responsibility to another role for the risk ownership).

- *Risk type*: should state which is the root-cause of the risk, e.g. human, operational, IT, supply chain, regulatory, etc.
- *Risk impact*: estimations, if possible, based on impact from known incidents, but also from impact statements from known vulnerabilities, as mentioned in vulnerability platforms (e.g. CWE and CVE (MITRE VWE; MITRE CVE)). Qualitative classification can also be used, e.g. through risk impact levels low, medium, high.
- *Risk probability*: can be quantitative or qualitative, derived accordingly from past incidents, audit reports, perceptions about quality of implemented protection, or available information from vulnerabilities (see *exploitability* and *evidence of existence* above). As regards risk exposure of IT- assets, a framework does exist that helps deriving the probability of an exploitation of a vulnerability (see Exploit Prediction Scoring System – EPSS (FIRST EPSS)). The method is based on CVS (MITRE CVS) and CVSS (FITST CVSS) (see also *Exploitability* of vulnerabilities).
- *Risk level*: the risk level can be evaluated by using various methods, such as quantitative, qualitative, semi-quantitative, statistical analysis on available historical data, etc. (Risk Analysis Methods) It is important to compare the assessed (*residual risk*) and the acceptable risks level defined for the involved asset relevant asset (see *Acceptable risk level)*. Given the available information in this risk assessment use case, the product to evaluate risk level can be modified to [(exploitability of vulnerability) x (impact on asset)]. This product is a good approximation of the risk level, irrespectively of which risk evaluation method has been chosen (i.e. *inherent* or *residual risks*).
- *Risk mitigation strategy*: the risk mitigation strategy can be selected as one of the options: treatment, avoidance, reduction, acceptance or transfer.

As regards **required resources**, this type of risk assessment requires a significant amount of work, whereas the project team (ca. four-five members), needs to have information security/cybersecurity skills. Hence, for a medium-sized enterprise, a risk assessment of – for example an application server implementing a few business functions – may require ca. two months. This would include an assessment of assets, vulnerabilities, and controls with the support of the IT-team.

Threat- and Asset-Based, Residual Risk Assessment

The consideration of threat information in risk assessment is not new. How-

ever, the proliferation of CTI has propelled the use of threat information and positively affected quality of risk assessment: the availability of detailed information about the whereabouts of threats is a strong tool that provides qualitative input to various parameters of vulnerabilities (i.e., exploitability) and risks (i.e., impact, probability). In the presented material (see Figure 1), threat related information consists of the entities: *Threat*, *Threat Agent* and *Attack Vector*.

CTI related information can be found en masse for IT-related assets, both as OSINT and restricted information. For other types of assets (i.e. information related, business, financial. etc.), some catalogues with more generic threats and weaknesses can be used (ISO 27005:2022). Open sources about impact and exploitation methods are far scarcer. Nonetheless, threat information and CTI is a game-changer in risk assessment.

As a matter of fact, there are cases where risk assessment is not performed based on assets, but rather **threat-oriented** (Dekker, M., Alevizos, L. 2023). Threat or Threat-Agent-oriented approaches do exist, whereas attention is not directly paid to assets and impact, but rather to threat-agent motivation, capability, skills and resources. Such approaches are rather applied in critical/strategic areas where successful attacks may lead to incidents in critical services and thus impact human life and national security.

Figure 5 shows the entities involved in this type of risk assessment and their relationships.

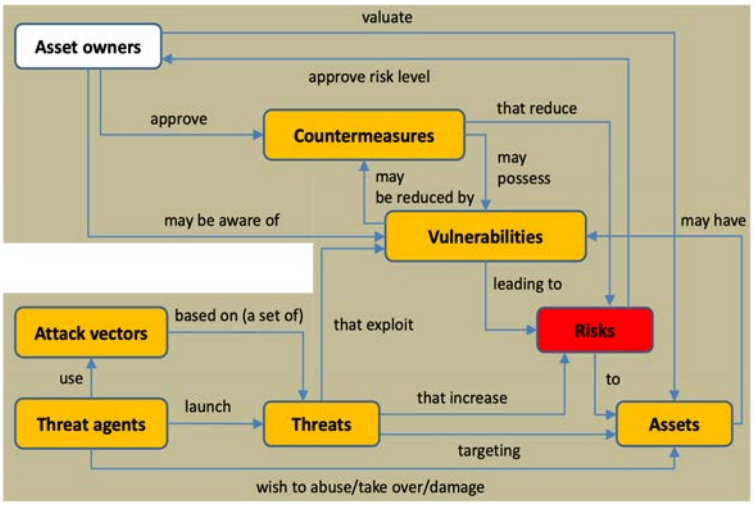

Fig. 5: Threat- and asset-based residual risk assessment

The contents of the entities: *Asset owners, Assets, Vulnerabilities, Countermeasures* and *Risks* will remain largely as in the *Asset-based, residual risk as-*

sessment (see the previous section). Should any changes/amendments in the content of these entities be necessary, they will be mentioned in the descriptions below.

As regards Threats:

- *Threat ID*: a unique ID. If an existing threat catalogue is used, threat IDs could be reused (e.g. ISO 27005 (ISO27005:2022), MITRE tactics (MITRE Enterprise Techniques), ENISA Threat Taxonomy (Marinos, L. 2016)).
- *Threat type*: indicates the domain from which the threat emanates, e.g. natural, environmental, failures, outages, technical, human, legal, etc. Such categories can be found in existing catalogues (same as in *ThreatID* above).
- *Threat details*: provides a description of the threat including: root cause, root phenomena, exploitation method, exploited weaknesses (vulnerabilities), used techniques, used artefacts, threat agent, etc.).
- *Ease of detectability*: indicates the time-span needed to detect the threat and the level of "stealthiness" of the threat (e.g. nature of exploitation traces).
- *Likelihood of appearance*: likelihood of threat appearance depends on various parameters, such as historical data and statistics, threat agent capabilities, motives and resources, efficiency/strength of countermeasure, *exploitability* of vulnerabilities, organizational readiness/procedures (i.e. organizational countermeasures) (Corona-Fraga, P. et al. 2026).
- *Known threat impact*: Impact statements for a threat are derived from the impact of weaknesses and the projection to the various areas of an organization. Main source for threat impact can emerge through cross reference of technical impact to business areas/missions (MITRE CWE Modeling the Environment). Herby, existing scorecards can be of value (MITRE CWRAF).
- *Known incidents*: enlists known incidents that have used this threat, with reference to attack vectors from successful attacks within which this threat has been deployed. A lot of information can be found for IT-threats. For other threat types of threat information (physical, human, processes, operations, etc.), the available threat information is significantly less (Physical Security Threats).

As regards Threat agents:

- *Threat agent name*: a name of the threat agent.
- *Access mode:* describes the access method of the agent for the initiation

of attack (s), e.g. internal, external, contractor infrastructure/access, supply-chain, etc.
- *Motivation*: indicates the assessed/assumed motivation of the threat agent, e.g. state-sponsored, activist, monetization, revenge, curiosity, etc.
- *Capability level*: describes the capabilities of the threat agent i.e. description of technical skills, sophistication, etc. Existing six tier levels can be used (Tse, A. 2022).
- *Resources/Means*: provides a description of used artefacts (i.e. threats, software), financial means, state support, infrastructures, etc.
- *Operation detectability*: description of the ease (or difficulty) to detect the threat agent actions, e.g. taken from military operations: overt, covert, clandestine (DeVine, M., E. 2022).
- *Known incidents*: list of attack vectors (campaigns (MITRE ATT&CK Campaigns)) attributed to the threat agent.

As regards Attack vectors:
- *Attack vector name*: name of the attack vector.
- *Attack vector description*: indicates the phases of the attack, the artefacts used (e.g., threats and tools), as well as sector and infrastructure. Could be eventually mapped to kill-chain phases.
- *Kown attacks*: provides known attacks that were based on this attack vector.
- *As regards Risks – Amendment*
- *Risk impact*: depends also on *Threats*, in particular *Known threat impact*. Which, in turn are based on *Exploitability* of exploited vulnerability.
- *Risk probability*: *Likelihood of appearance* flows in the estimation of risk probability, together with all other parameters. It supports the estimation, by also taking into account the capability level of the corresponding *threat agent*.
- *Risk level*: Given the available information in this risk assessment use-case, the product to evaluate risk level can be modified to [(likelihood of appearance of Threat) x (exploitability of vulnerability) x (impact on asset)].

Concluding this section, one should note the rather high complexity of threat information for IT-related assets, but also the large amount of available detailed with comprehensive coverage. These two facts make the performance of threat-based assessments ideal for the information technology groups of organizations. This is the reason why this practice is followed by organizations with medium-large infrastructures and IT-security teams, with the latter often

become the risk-owners of the risks related to IT-assets. On the other hand, this high complexity is difficult to master within organizations with small IT-teams and outsourced IT-infrastructure. For those, the selection of *asset-based, residual risk assessment* will be eventually the risk assessment use-case of choice.

Another noticeable element of this use-case is, that if the risk assessment concentrates on other than detailed IT-assets, e.g. business processes, financial, human, physical, infrastructures, etc., then the availability of data to perform threat-based assessment is significantly lower and of less detail. In that case, a tailoring of the threat part of this assessment use-case will be necessary. This will reduce complexity according to the completeness and quality of available threat information for the assets in focus.

As regards **required resources**, this type of risk assessment can be very laborious, requiring a team with skills in threat intelligence and threat hunting. Few examples of such projects that fell in the attention of the author were all in large enterprises with own threat analysis teams. For a major server infrastructure, for example, the majority of the threat analysis team (ca. five to seven experts), have been working on a threat-based risk assessment for ca. four to five months. More "light" versions of such assessments (simpler infrastructure and less threat analysts) may be around four months, occupying a team of ca. five persons (one threat analyst).

4. Risk Assessment of Layered, Interdependent Complex Systems

Given the complexity of business and services infrastructures, risk assessments need to follow a structured way of performance according to the various levels of operations, from business processes, to services, to infrastructures, to applications and data. The differentiating nature of those levels, but also existing interdependencies need to be taken into account when performing risk assessments. This creates the necessary coherence towards achieving quality results in risk assessment. Capturing these interdependencies is decisive for assessing and managing risks through the (up-stream and down-stream) orchestration of implemented countermeasures. In this section we present decomposition techniques used to identify interdependencies and we provide a mapping to the previously discussed risk assessment use-cases.

The task to be performed resembles the decomposition method of an enterprise architecture (TOGAF Standard 2022). Along with some various approaches found (Sherwood, J. et al. 2005; Component Driven Approach), in the following discussion a component driven approach (Component Driven Approach) is presented.

The component analysis follows a top-down decomposition, divided into

different levels following a logic leading from the generic to the specific. In particular:

- Level 1: Critical business processes, information and services
- Level 2: Architecture components, information flows, trust boundaries, exit/entry points
- Level 3: People, Access, and Personally Identifiable Information (PII)
- Level 3: Computing hardware and devices
- Level 4: Network infrastructure
- Level 5: Operating systems, applications, functions and APIs
- Level 6: Data, data storage, data flows
- Level 7: Vendors, contractors, external supply-chain

The above enlisted levels are subject to scoping. As such, depending on the width and breadth of the focus of risk assessments, they can be selected according to the needs, sector and risk levels, usually derived after the risk assessment of level 1. To this extend, the list above is indicative and the decomposition levels may vary.

Depending on the decomposition level, a different risk assessment use-case may be appropriate. In the table below (see ANNEX, Table 2), some proposals for the use of risk assessment use-cases are presented per level. The table presents also the enrolment of the entities in the consolidation assessment results between the various levels (both top-down and bottom-up).

5. Concluding Remarks

Concluding this chapter, we would like to provide some conclusions from own experience in risk assessment. Here comes the **top 5**:

1. Regulatory compliance is the beginning, **not** the destination of a risk assessment journey. State this explicitly to any risk owner who cares only about tick-box exercises.
2. Careful risk assessments unveil – almost always – **new assets**, that the involved company representatives never thought of. If your project did not experience this, be aware about the thoroughness and efficiency of your method and/or involved individuals.
3. Manage the expectations of the involved companies regarding upfront discussions about cost/benefit ratios of a risk assessment project. If the work is properly scoped and addresses the right level of detail, it is always clear that it is worth the effort.

4. Follow the principle "**less is more**". Do not overload a project with wide scope or detail depth. Adapt to the needs of the organization, to the criticality of the business, to available skills and the follow-up costs of the risk assessment. Sometimes it is better to do something leaner and people can take it further, than something complex that none is in the position to pursue further.
5. Be aware that almost any risk assessment project may cause **internal tensions** in an organization: not every executive/manager is happy to hear that their area bears risks that need to be managed. They might be concerned about reputation, generation of costs and generally unrest in the management floor. Try to form your assessment teams in such a way, as to minimize frictions.

References

ISO 27005:2022, *Information security, cybersecurity and privacy protection – Guidance on managing information security risks*. 4th Edition, 2022. International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).

Dekker, M., Alevizos, L. 2023. *A threat-intelligence driven methodology to incorporate uncertainty in cyber risk analysis and enhance decision-making*. Security and Privacy published by John Wiley & Sons Ltd. Online Library. Available online at https://knowledge.lancashire.ac.uk/id/eprint/45776/1/Security%20and%20Privacy%20-%202023%20-%20Dekker%20-%20A%20threat%E2%80%90intelligence%20driven%20methodology%20to%20incorporate%20uncertainty%20in%20cyber%20risk.pdf, (Last accessed: October 17, 2026).

Vulnerability 2026. Wiktionary, The free dictionary. Last edited on 13 October 2026. Available online at https://en.wiktionary.org/wiki/vulnerability (Last accessed: October 17, 2026).

Threat. *Glossary*. NIST Computer Security Resource Center (CSRC). Available online at https://csrc.nist.gov/glossary/term/threat (Last accessed: October 17, 2026).

Hutchins, E., Cloppert, M., Amin, R., *Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains*, White Paper, Lockheed Martin Corporation. Available online at https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf (Lat accessed October 17, 2026).

MITRE ATT&CK Campaigns. The MITRE Corporation. Available online at https://attack.mitre.org/campaigns/ (Last accessed: October 17, 2026).

CIA Triad. Wikipedia the free encyclopedia. Last edited on 15 September 2026. Available online at https://en.wikipedia.org/wiki/CIA_triad (Last accessed: October 17, 2026).

RACI Matrix. Wikipedia the free encyclopedia. Last edited on 23 July 2026. Available online at https://en.wikipedia.org/wiki/Responsibility_assignment_matrix (Last accessed: October 17, 2026).

Marinos, L. 2016. ENISA *Threat Taxonomy: A tool for structuring threat information*. European Union Agency for Cybersecurity. Published ENISA Report. Available online at https://www.researchgate.net/publication/387861057_ENISA_Threat_Taxonomy_A_tool_for_structuring_threat_information_ENISA_Threat_Taxonomy_About_ENISA_Author_ENISA_Threat_Taxonomy (Last accessed: October 17, 2026).

ISO 15408-1:2026. *Information security, cybersecurity and privacy protection – Evaluation criteria for IT security Part 1: Introduction and general model*. 5th Edition, 2026. International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).

Andrukiewicz, E., Atallah, E., Bartels, C., Dörr, C., Marinos, L., Michota, A., Mongay Batalla, J., Nguyen, K., Payen, P., Polemi, N., Preneel, B., Quemard,

J-P., Schaumüller-Bichl, I., Stütz, G., Ward, J. 2021. *METHODOLOGY FOR SECTORAL CYBERSECURITY ASSESSMENTS – EU Cybersecurity Certification Framework.* European Union Agency for Cybersecurity. Published ENISA Report. Available online at https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20SCSA%20Methodology.pdf (Last accessed: October 17, 2026).

Risk Matrix. Wikipedia the free encyclopedia. Last edited on 29 August 2026. Available online at https://en.wikipedia.org/wiki/Risk_matrix (Last accessed: October 17, 2026).

ISO 27001:2022. *Information security, cybersecurity and privacy protection – Information security management systems – Requirements*. 3rd Edition, 2022. International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).

ISO 31000:2018. *Risk management – Guidelines*. 2nd Edition, 2018. International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC).

MITRE CWE. CWE – *Common Weakness Enumeration*, 2006-2026. MITRE Corporation. Available online at https://cwe.mitre.org, (Last accessed: October 17, 2026).

MITRE CVE. CVE – Common Vulnerabilities and Exposures, 1999-2026. Available online at https://www.cve.org, (Last accessed: October 17, 2026).

FIRST CVSS. *CVSS – Common Vulnerability Scoring System*, 2015-2026. 4th Version. Forum of Incident Response and Security Teams, Inc. Available online at https://www.first.org/cvss/, (Last accessed: October 17, 2026).

Indicator of Compromise (IoC). Wikipedia the free encyclopedia. Last edited on 11 August 2026. Available online at https://en.wikipedia.org/wiki/Indicator_of_compromise, (Last accessed: October 17, 2026).

Lenaerts-Bergmans, B. 2025. *INDICATORS OF ATTACK (IOAS).* CrowdStrike. Available online at https://www.crowdstrike.com/en-us/cybersecurity-101/threat-intelligence/indicators-of-attack-ioa/, (Last accessed: October 17, 2026).

FIRST EPSS. *Exploit Prediction Scoring System (EPSS)*, 2015 – 2026. Forum of Incident Response and Security Teams, Inc. Available online at https://www.first.org/epss/, (Last accessed: October 17, 2026).

Risk Analysis Methods. Section "*Types of Risk Analysis Methods*", 2026. MetricStream. Available online at https://www.metricstream.com/learn/risk-analysis-types-methods.html#section-7, (Last accessed: October 17, 2026).

MITRE Enterprise Techniques. MITRE Corporation. Available online at https://attack.mitre.org/techniques/enterprise/, (Last accessed: October 17, 2026).

Corona-Fraga, P., Díaz-Rodriguez, V., Niebla-Zatarain, J. M., 2026. *Measuring Risk in Cybersecurity via Likelihood*. In *Proceedings of First Summer School on Artificial Intelligence in Cybersecurity.* Available online at https://www.mdpi.com/2673-4591/123/1/39, (Last accessed: October 17, 2026).

MITRE CWE Modeling the Environment. CWE – *Common Weakness Enumeration*, 2006-2026. Modeling the Environment: Business Domains, Technology Groups, Archetypes, and Vignettes. MITRE Corporation. Available online at https://cwe.mitre.org/cwraf/modeling_the_env.html, (Last accessed: October 17, 2026).

MITRE CWRAF 2023. *Common Weakness Risk Analysis Framework*,. MITRE Corporation. Available online at https://cwe.mitre.org/cwraf/, (Last accessed: October 17, 2026).

Physical Security Threats. *Data Sources for Physical Security Threat Assessments: The Complete Guide*. Base Operations, USA. Available online at https://www.baseoperations.com/article/data-sources-for-physical-security-threat-assessments, (Last accessed: October 17, 2026).

Tse, A. 2022. *The Different Types Of Threat Actors And Their Motivations*. Medium Human stories & ideas. Available online at https://medium.com/codex/the-different-types-of-threat-actors-ce04852599e2, (Last accessed: October 17, 2026).

DeVine, M., E. 2022. *Covert Action and Clandestine Activities of the Intelligence Community: Selected Definitions*. Library of Congress, USA. Available online at https://www.congress.gov/crs-product/R45175, (Last accessed: October 17, 2026).

TOGAF Standard 2022. *TOGAF – The Open Group Architecture Framework*. 10th Edition. The Open Group. Available at https://www.opengroup.org/togaf, (Last accessed: October 17, 2026).

Sherwood, J., Clark, A., Lynas, D. 2005. *Enterprise Security Architecture: A Business-Driven Approach*". CRC Press. Found also as SABSA – *Sherwood Applied Business Security Architecture*. SABSA Institute C.I.C. Complete SABSA reference is available online at https://sabsa.org/sabsa-executive-summary/, (Last accessed: October 17, 2026).

Component Driven Approach. *Component driven risk management methods*. National Cyber Security Centre (NCSC), UK. Available online at https://www.ncsc.gov.uk/collection/risk-management/introducing-system-and-component-driven-risk-management-approaches, (Last accessed: October 17, 2026).

ANNEX

Table 1.: Full set of risk assessment entities, their relationships and content

Entity	Possible content	Its relationships	Transitive content
Asset	*Asset name* *Asset type*: (tangible, intangible, human, physical, financial, etc.) *Asset value*: (possibly qualitative) *Acceptable risk level:* indicates the level of acceptable risk exposure for the asset. It is mostly defined qualitatively (i.e. medium, low, very low). The acceptable risk level is indicative for the level of protection requirements fulfillment (see below). *Protection requirements*: required/ implemented protection levels (e.g. via CIA Triad: Confidentiality, Integrity, Availability (CIA Triad))	is evaluated by asset owner may have vulnerabilities can be abused by threat agent targeted by threat Asset exposure to risks	Value set by *Asset owner* *Asset vulnerabilities* *Treat agent name(s)* *Threats targeting the asset* *Risks to the asset*
Asset owner	*Name of the asset owner*: physical person, organization unit, etc. *Assumed role for asset(s)*: (e.g. based on RACI – responsible, accountable, concerned, informed (RACI Matrix))	valuates owned asset approves risk level for owned asset approved countermeasures is aware of vulnerabilities	Sets *Asset value* *Risk level identification* (for assets with accountability) *List of approved countermeasures* *List of assessed vulnerabilities*

Entity	Possible content	Its relationships	Transitive content
Vulnerability	*Name of vulnerability* *Type of weakness*: technical, organizational, other *Exploitability*: (e.g. exploitation methods, level of difficulty to exploit, etc.) *Evidence of existence*: scans, previous incidents, indicators, etc.	awareness of asset owner affected asset(s) reduced through countermeasure(s) caused by countermeasure list of related risks exploited through threat	*Asset owner(s)* affected *asset(s)* Related *countermeasure* countermeasure has *vulnerability* exploitation of vulnerability materializes *risk*(s) vulnerability can be exploited by *threat(s)*
Threat	*Threat ID* *Threat type*: (natural, environmental, failures, outages, nefarious activities/abuse, legal, etc. (Marinos, L. 2016)) *Threat details*: (description of threat, method, techniques, etc.) *Ease of detectability* *Likelihood of appearance* *Known threat impact*: (e.g. qualitative, including CIA triad, etc.) *Known incidents*	exploits vulnerability(ies) launched by threat agent may increase risk(s) level is part of attack vector(s) targets asset(s)	*Vulnerability(ies)* exploited by the threat *Threat agent* uses this threat Theat increases *risk(s)* *Attack vector* List of affected *assets*

Entity	Possible content	Its relationships	Transitive content
Risk	*Risk ID* *Risk title* *Risk owner*: (the role who is accountable for the assessment and treatment of a risk. It can be the asset owner or a person mandated to manage the risk on behalf of the asset owner[1]) *Risk type*: (partially derived by the nature of the related assets, i.e. human, operational, IT, supply chain, regulatory, etc.) *Risk impact*: (qualitative, eventually mapped to CIA triad) *Risk probability*: (derived from threat likelihood, threat agent capability, exploitability of vulnerability and frequency of incidents) *Risk level*: (qualitative indication for the level of acceptable risk) *Risk mitigation strategy*: (one of: avoidance, reduction, acceptance, transfer)	risk(s) impacts asset(s) risk occurs via materialization of threat risk(s) reduced through countermeasure(s) vulnerability may pose risk level of risk has been approved by asset owner	*Asset(s)* impacted by the occurrence of the risk *Threat(s)* caused the incident *Countermeasure(s)* reducing threat exposure *Vulnerability*(ies) whose exploitation led to the risk *Asset owner*[2] who approved the level of risk level of this risk
Countermeasure	*CountermesureID* *Countermeasure type*: (technical, organizational, physical, etc.) *Responsible role*: (role that carries the responsibility for the implementation and management of the countermeasure) *Protection objectives*: (e.g. CIA mapping)	approved by asset owner reduces vulnerability mitigates/reduces risk(s) may possess vulnerability	*Asset owner* (i.e. via RACI role) *Vulnerability(ies)* whose exploitation is reduced through the countermeasure *Risk*(s) reduced by this countermeasure Possible *vulnerability* of the countermeasure itself.

[1] Depending on the nature of the asset, an asset owner can be hierarchically a higher or lower role than the risk owner (see also section "Modus operandi / Use-Cases of risk assessment"). This really depends on the hierarchical distribution of responsibilities in the organization.

[2] Here the reference is to the owner of the top-level asset, potentially at business level. This discrimination is relevant when through the necessary asset decomposition, lower-level assets are assigned to lower roles in the hierarchy. Those roles should not carry the full accountability for the top-level asset, as they might have been mandated to manage (or be responsible for) this asset.

Entity	**Possible content**	**Its relationships**	**Transitive content**
Threat agent	*Threat agent name* *Access mode:* (e.g. internal, external, contractor, supply-chain, etc.) *Motivation*: (e.g. state-sponsored, activist, monetization, revenge, curiosity, etc.) *Capability level*: (i.e. description of technical skills, sophistication, etc.) *Resources/Means*: (descriptive, including e.g. tools, financial means, state support, infrastructures, etc.) *Operation detectability*: (descriptive, e.g. ease of detection of actions, mode of operation – overt, covert, clandestine) *Known incidents*	launches threats uses attack vectors targets asset(s)	List of known *threats* from threat agent Threat agent uses *attack vector(s)* *Asset(s)* targeted by threat agent
Attack vector (used as synonym to Campaign)	*Attack vector name* *Attack vector description*: (e.g. including techniques and tools, eventually mapped to kill-chain phases) *Kown attacks*	based on (a set of threats) used by threat agents	List of *threats* that are part of the attack vector The attack vector is used by *threat agents*

Table 2.: Various decomposition levels for risk assessment of layered, interdependent complex systems

Decomposition level	Proposed risk assessment use-case	Decomposition guidance (to more detailed level)	Consolidation guidance (to more generic level)
L1 Critical business processes, information and services	*Rudimentary risk assessment*	*Rudimentary asset groups* (processes, services, information), prioritized according to their *risk levels* are passed to next decomposition	No further generic assessment assumed
L2 Architecture components, information flows, trust boundaries, exit/entry points	*Basic, asset-based assessment*	*Asset* list prioritized according to their *risk levels* are passed to next decomposition	Consolidation based on *Risks* of more detailed levels Consolidation with *Assets / Asset groups* of the next generic level
L3 People, Access, and Personally Identifiable Information (PII)	*Basic, asset-based assessment* *or (for higher criticality)* *Asset-based, residual risk assessment*	*Asset list* prioritized according to their risk levels are passed to next decomposition *or (for higher criticality)* *Assets*, *vulnerabilities* and *countermeasures* list are passed to next decomposition	Consolidation with *Risks* and *Assets* of next generic level *or (for higher criticality)* Consolidation with *assets, vulnerabilities and countermeasures* with the *asset, vulnerabilities and countermeasures list* of next generic level
L4 Network infrastructure	*Asset-based, residual risk assessment* *or (for higher criticality)* *Threat- and asset-based residual risk assessment*	*Assets*, *vulnerabilities* and *countermeasures* list are passed to next decomposition *or (for higher criticality)* *Assets*, *vulnerabilities, threats, threat agents, attack vectors* and *countermeasures* list are passed to next decomposition	Consolidation with *Risks* and *Assets* of next generic level
L5 Operating systems, applications, functions and APIs	*Asset-based, residual risk assessment* *or (for higher criticality)* *Threat- and asset-based residual risk assessment*	No further detailed assessment assumed	Consolidation with *Risks, Assets*, *Vulnerabilities* and *countermeasures* list of next generic level *or (for higher criticality)* Consolidation with *Risks, Assets*, *vulnerabilities, threats, threat agents, attack vectors* and *countermeasures* list of next generic level

Decomposition level	Proposed risk assessment use-case	Decomposition guidance (to more detailed level)	Consolidation guidance (to more generic level)
L6 Data, data storage, data flows	*Asset-based, residual risk assessment*	No further detailed assessment assumed	Consolidation of *Risks* with the *Risks* of the next generic level Consolidation of *Assets* with the *Assets/Asset groups* of next generic level
L7 Vendors, contractors, external supply-chain	*Basic, asset-based assessment* *or (for higher criticality)* *Asset-based, residual risk assessment*	No further detailed assessment assumed	Consolidation of *Risks* with the *Risks* of the next generic level Consolidation of *Assets* with the *Assets* of next generic level *or (for higher criticality)* Consolidation with *Risks, Assets*, *Vulnerabilities* and *countermeasures* list of the next generic level

REDEFINING CYBERSECURITY GOVERNANCE AGAINST ADAPTIVE THREATS: FROM CLASSICAL RISK ANALYSIS TO AUTONOMOUS CYBER-IMMUNITY

Carmelo Dimauro
LHC – Luxembourg House of Cybersecurity
National Cybersecurity Competence Center, Luxembourg

Abstract
This chapter examines the structural limitations of traditional risk quantification frameworks within adaptive digital environments, focusing specifically on the governance implications of Agentic Artificial Intelligence. For decades, corporate risk management has relied on the Kaplan-Garrick framework to structure scenarios, probabilities, and impacts. While historical data remains essential for understanding vulnerability prevalence, tracking control effectiveness, and mapping institutional exposure, its utility as a sole basis for predicting future cyber incidents decreases when encountering self-mutating malware and autonomous, goal-oriented adversaries. This shift transitions the cyber landscape into an environment characterised by Knightian uncertainty, where the assumptions underlying traditional models, stable scenarios, meaningful probability estimates, and bounded system behaviour, become increasingly difficult to maintain. As manual intervention proves too slow to counter machine-speed strikes, cybersecurity governance must evolve to manage a continuous, strategic interaction between AI attackers and AI defenders. Consequently, organisations must supplement traditional risk prevention with a barbell strategy that protects a conservative operational core while allowing flexible, autonomous defence at the perimeter. This approach leads to a model of cyber-immunity, where systems mimic biological organisms capable of self-healing. Crucially, the paper addresses the autoimmune problem of autonomous cyber-defence, where over-aggressive defensive AI inflicts systemic disruptions through miscalibrated quarantine cascades, and redefines Governance, Risk, and Compliance (GRC) as an architectural framework designed to constrain, audit, and establish explicit autonomy boundaries around autonomous self-healing systems.

1. Introduction

The Borrowed Foundations of Corporate Security

The architectural foundations of modern corporate governance were designed primarily for predictable physical environments. When early information systems integrated into the fabric of enterprise operations, executive boards naturally borrowed risk management frameworks from long-established disciplines like civil engineering, physical manufacturing, and actuarial science. These legacy models operate on a deductive premise, suggesting that by ana-

lysing the frequency of past incidents, organisations can map the mathematical probability and financial impact of future failures. Risk, under this view, is managed as a closed system governed by stable statistical distributions, much like the predictable, natural deterioration of mechanical infrastructure.

The Machine-Speed Reality

The rapid integration of Artificial Intelligence has challenged this foundational assumption, exposing an operational mismatch between modern threats and human response times. The digital ecosystem is no longer a passive environment subjected to random wear. Instead, it has become a fluid environment characterised by autonomous Agentic AI, self-mutating polymorphic malware, and automated zero-day exploitation tools. When offensive assets can independently discover vulnerabilities, alter their own digital signatures, and execute complex lateral strategies at machine speed, historical data loses its sufficiency as a sole predictive tool. The corporate world has unknowingly crossed a critical boundary, moving away from a realm of calculable risk and into a state of absolute, unmeasurable uncertainty.

Structure of the Inquiry

This paper investigates how the meaning of cybersecurity risk, and its subsequent governance, changes when both the attacker and the defender become adaptive, autonomous agents. The text charts a clear logical journey. It begins with the dismantling of traditional risk metrics and moves systematically through the realities of algorithmic warfare. Rather than treating advanced technical solutions as a simple upgrade to existing firewalls, this paper outlines a structural paradigm shift. By connecting economic theory, game mechanics, and biological analogies, it establishes a new mandate for executive *leadersh*ip, redefining governance not as a tool to prevent attacks, but as the framework required to manage autonomous survival and control the inherent dangers of digital self-healing.

2. The Foundations of Classical Cyber-Risk and the Fallacy of Predictability

The Deductive Paradigm

For decades, industrial, corporate, and digital risk management has operated under the assumption that the future can be reliably mapped using the coordinates of the past. When organisations sought to protect their physical infrastructure, financial portfolios, or early information technology systems, they looked to disciplines with long, established traditions. From these fields, modern cybersecurity governance inherited a quantitative, deductive para-

digm, building an entire architecture of compliance, audits, and strategic decision-making around the assumption that risks could be systematically identified, measured, and mitigated. At its core, this traditional approach treats the digital ecosystem as a closed, predictable market, assuming that cyber threats operate within stable distributions that repeat over time.

The Kaplan-Garrick Triplet

To govern this environment, corporate boards and risk managers frequently rely on the foundational conceptual architecture established by industrial engineers Stanley Kaplan and B. John Garrick. In their seminal formulation, Kaplan and Garrick (1981) provided a structured method for representing risk as a set of triplets, explicitly addressing the interplay between scenarios (S_i), the likelihood or probability of those scenarios (L_i), and their resulting consequences (C_i). Crucially, their model was never a simple exercise in historical data regression; they explicitly accounted for uncertainty, the problem of completeness, and the boundaries of risk acceptability, treating probability as a rigorous representation of a state of knowledge rather than a static historical frequency. Early adoptions of total risk management frameworks, as outlined by Haimes (1991), expanded on this logic by establishing rigorous protocols to systematically analyse multivariable engineering complexities under well-defined operational assumptions.

The difficulty introduced by modern, adaptive cyber threats is not that the Kaplan-Garrick triplet is conceptually invalid, but rather that each structural component becomes progressively more difficult to establish with analytical confidence. When the system, the adversary, and the defensive response are all adaptive, the inputs required to stabilise the triplet are thrown into constant flux.

First, the question of scenario identification, what can happen, becomes increasingly difficult to map when attack paths are generated dynamically by autonomous agents rather than human actors following pre-scripted methodologies. Second, establishing the likelihood of an occurrence becomes problematic because the adversary adapts strategically and reactively to the defensive perimeter, invalidating stationary statistical baselines. Third, quantifying the expected consequences becomes highly complex because the defensive responses themselves can alter system behaviour in real time, creating secondary operational impacts that are difficult to bound a priori. The framework remains an essential conceptual tool, but its operational execution is strained by the fluid dynamics of algorithmic environments.

Residual-Risk Acceptance

By calculating these variables, organisations arrive at a number or a qualitative rating, such as a high, medium, or low classification. This metric

allows them to engage in the conventional logic of residual-risk acceptance, drawing heavily on probability and statistics within classical queueing theory and stationary distributions used to quantify computer science reliability metrics (Trivedi 2002). In this legacy corporate mindset, executives look at the calculated probability and the expected impact, deploy a series of static controls like firewalls or password policies to reduce the risk, and then accept whatever margin of danger remains. This deductive loop functions effectively in environments where historical data remains representative of future frequencies. However, the operational implementation of this framework often relies on the assumption of a static baseline. As we observe the integration of hyper-complex cloud ecosystems and interconnected digital supply chains, the boundaries of these static models are reached. The challenge is not that classical risk analysis is fundamentally flawed, but rather that its structural assumptions become increasingly insufficient when the threat landscape transitions from a passive environment to a dynamically adaptive one.

3. The Shift to Knightian Uncertainty

Risk versus Deep Uncertainty

As we step out of the predictable world of classical risk management, we immediately encounter a conceptual boundary. Traditional frameworks assume that even if we do not know when a cyber-incident will occur, we can at least calculate the mathematical odds of it happening. However, the modern digital landscape, accelerated by the rapid integration of Artificial Intelligence, challenges this assumption. To understand why our mathematical models are failing, we must look to an essential distinction made over a century ago by the economist Frank Knight.

In his seminal work, Knight (1921) drew a sharp line between two ideas that are frequently confused in corporate boardrooms: calculable risk and genuine uncertainty. According to Knight (1921), risk describes a situation where the future is unknown, but we can accurately calculate the probabilities of different outcomes. This is the domain of Kaplan and Garrick, relying on a stable environment where past observations create a reliable map of future frequencies. Uncertainty, or what academics call Knightian uncertainty, describes a completely different realm. It represents situations where the future outcomes are fundamentally unquantifiable. You cannot calculate the probabilities because you simply do not have the data, the system rules are constantly changing, or the event has never occurred before.

The Failure of Stochastic Inference

The modern cyber ecosystem, driven by autonomous digital assets and complex cloud structures, has crossed the boundary from the traditional definition of risk into the territory of absolute uncertainty. Traditional stochastic inference, the practice of using historical data to predict future failures, breaks down here. As established in the macroeconomic framework of cybersecurity economics by Kianpour, Kowalski, and Øverby (2021) [MDPI Sustainability Journal], using past incident frequencies to predict future digital exposure is fundamentally limited within fluid institutional environments. This limitation arises because cybersecurity does not contend with passive decay, but with an intelligent, adaptive adversary.

When AI is introduced into the mix, this adversary can continuously mutate its techniques. If a threat actor uses automated tools to generate entirely new, self-evolving attack vectors, historical databases become less sufficient as a basis for predicting adaptive behaviour. Relying on last year's data to predict tomorrow's AI-driven breach leaves an organisation blind to systemic shocks and what Taleb (2012) characterises as Black Swan events, which represent highly consequential, unpredictable anomalies that lie entirely outside the realm of ordinary expectations. Consequently, the core challenge of modern cybersecurity governance is no longer about refining mathematical equations to assess measurable probabilities. The challenge is learning how to govern under epistemic uncertainty, a state where the very rules of the system are in constant, unpredictable flux.

The Three Vectors of Recursive Uncertainty

This transition into epistemic uncertainty can be deconstructed into three distinct, interconnected vectors that modern governance frameworks must confront. The first vector is uncertainty about the event: will a highly consequential cyber incident occur, and through what entry point will it manifest? The second vector is uncertainty about the adversary: once inside a perimeter, how will an intelligent, strategically adaptive attacker behave when encountering architectural resistance? These first two dimensions are relatively familiar to risk practitioners operating under deep uncertainty.

Parallel to this challenge, the integration of autonomous systems introduces a critical, third dimension: uncertainty about the defence. When an organisation delegates real-time operational response to an autonomous, algorithmic defender, it can no longer predict with certainty how that defensive mechanism will interpret internal systemic anomalies or execute its remediation protocols. This third vector introduces a state of recursive uncertainty. The organisation is no longer simply uncertain about the external threat; it becomes structurally uncertain about the behaviour of the internal system deployed to manage

that threat. In this environment, uncertainty is no longer an exogenous variable driven solely by an outside adversary. It becomes an endogenous, recursive loop where the actions of the defender dynamically reshape the operating environment. It is at this precise theoretical intersection that the traditional concepts of risk control reach their conceptual limit, making the biological paradigm of the autoimmune problem an intellectual necessity. To govern a modern enterprise is to govern a system that is structurally exposed to the recursive, unpredictable feedback loops of its own automated shields.

The conceptual progression developed in this chapter can be summarised as a transition from classical cyber-risk towards an adaptive and ultimately autonomous paradigm. Figure 1 illustrates this evolution, highlighting the point at which the defence mechanism itself becomes an object of governance.

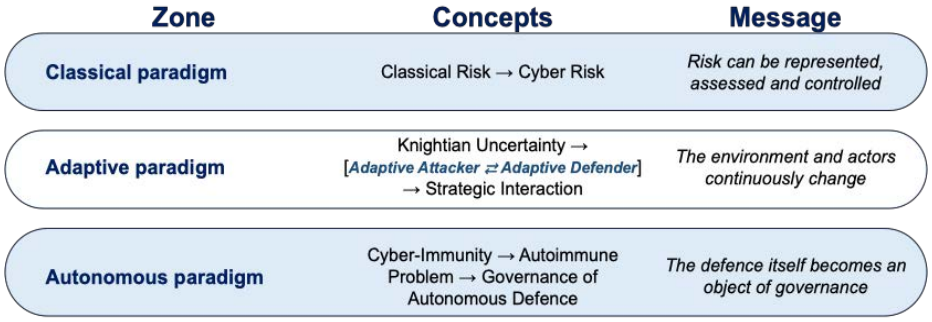

Fig. 1: Conceptual evolution of cybersecurity risk and governance in adaptive digital environments

4. The Advent of the Adaptive Attacker

The Automated Threat Landscape

To comprehend how Knightian uncertainty manifests in the digital real world, we must examine the broader governance implications of AI-assisted offensive systems. In the classical era of cybersecurity, malicious operations were constrained by human limitations, requiring manual exploration, chronological script writing, and human oversight. Today, the integration of Agentic AI and adaptive workflows has altered this landscape, allowing threat actors to automate significant portions of reconnaissance, vulnerability analysis, code generation, social engineering, and attack execution. This systemic evolution is documented by the European Union Agency for Cybersecurity (ENISA 2024), which notes that threat groups are leveraging machine learning paradigms to refine their targeted pipelines and scale infrastructural exploitation.

The operational reality of these automated threats has crossed from conceptual modelling into active telemetry. Empirical analyses of early automated Large Language Model orchestrations and the security vectors of Model Context Protocol configurations have shown that adversarial networks based on Agentic AI can systematically coordinate complex lateral expansion frameworks (Shapira 2025, Guo et al. 2025, Zhao et al. 2025). Furthermore, verified threat research has demonstrated that modern autonomous architectures possess the capability to independently execute automated vulnerability scans, perform sandbox escalation, and complete autonomous web hacking sequences far faster than manual security teams can react (Wiz Research 2026).

The Framework Alignment

This paradigm shift is reflected in recent institutional frameworks, notably the Cybersecurity Framework Profile for Artificial Intelligence, or Cyber AI Profile (NIST IR 8596), developed by the National Institute of Standards and Technology (NIST 2025). The NIST architecture explicitly splits the risk ecosystem into three distinct focus areas: securing the internal AI system components, conducting AI-enabled cyber defence, and thwarting AI-enabled cyberattacks. By formally isolating AI-enabled cyberattacks as a unique dimension of risk, standard-setting bodies recognise that adaptive offensive capabilities modify traditional defence frameworks. When an offensive system can independently alter its execution parameters, the stable scenarios required by legacy risk assessment methodologies change, leaving organisations exposed to automated campaigns that operate faster than human verification timelines.

Pillars of Evolving Exploitation

- **Hyper-Personalised Cognitive Social Engineering**: Agentic AI frameworks have altered the economics and scalability of targeted deception. Historically, social engineering campaigns frequently contained discernible indicators, such as grammatical anomalies or awkward phrasing. Advanced agentic architectures have removed these stylistic indicators, enabling threat actors to replicate internal corporate communication styles and generate contextualised, clear messages at scale. Combined with real-time deepfake audio and video synthesis, these tools reduce the effectiveness of traditional organisational awareness measures that rely heavily on employees recognising conventional linguistic or stylistic warning signs. Consequently, corporate security cultures can no longer treat human intuition as a primary line of perimeter defence, necessitating a pivot toward cryptographic authentication and automated structural validation.
- **Polymorphic Execution Fabrics:** The malicious payload itself has become adaptive and self-evolving. Modern offensive tools utilise polymorphic

mutation engines driven by Agentic AI that dynamically alter the software's underlying syntax during runtime while preserving its operational intent. By shifting its code structure mid-execution, the malware bypasses static signature verification and traditional heuristic detection methods.

When threats can reason, self-replicate, and coordinate campaigns at machine speed through Agentic AI, the operational execution of the Kaplan-Garrick framework faces distinct limitations. Scenario identification becomes exceptionally complex because the attack path is actively modified by an autonomous algorithm in response to defensive countermeasures. The probability of a breach is no longer a calculable statistic, but a shifting variable controlled by an intelligent, goal-oriented adversary.

5. The Symmetric Shield and the Game-Theoretic Duel

The Limits of Manual Defence

The realisation that offensive AI operates at machine speed forces an operational conclusion upon cybersecurity governance, as human defenders are too slow to counter automated threats in real time. When an agentic asset can find a vulnerability, generate an exploit, and compromise a network within minutes, relying on human analysts to detect an alert, hold an administrative meeting, and manually isolate a server fails to protect the infrastructure. To withstand an automated, adaptive threat, the defensive architecture must operate within the exact same compressed timeline. Organisations must deploy defensive AI tools. These automated defensive systems are tasked with continuously monitoring network traffic, predicting potential vectors, and executing incident response processes without waiting for human authorisation.

Strategic Interaction Dynamics

Consequently, the nature of cybersecurity undergoes a functional transformation. It ceases to be an exercise in static risk analysis or compliance-driven prevention, becoming instead a direct, real-time battle between two autonomous software agents: the AI attacker and the AI defender. This symmetric confrontation cannot be accurately described or managed using traditional corporate governance models. Because both the attacker and the defender are intelligent, reactive, and capable of modifying their strategies based on the actions of the opponent, this environment is better understood as a strategic interaction in which each side observes, responds to, and attempts to anticipate the behaviour of the other, drawing on early foundations established by von Neumann and Morgenstern (1944).

To govern this automated environment, frameworks must treat network states as complex discrete event systems where state changes are driven by instantaneous events rather than continuous parameters (Cassandras and Lafortune 2008). In a conventional risk model, the threat is something passive that happens to the organisation, much like a winter storm or a decaying component that does not change its nature because an enterprise has bought insurance. In a strategic environment, however, the threat is something that actively reacts to the organisation. Every defensive action taken by the enterprise may provide critical information to the attacker, who can then adapt their approach. The defender, in turn, observes this new behaviour and adapts again. Cyber-risk therefore becomes part of an ongoing interaction rather than a fixed property of the system. For governance boards, this realisation changes the definition of operational control. You can no longer audit a system based on whether a specific door is locked, because the adversary is actively studying the nature of the lock, and the automated shield is constantly altering the digital environment to disrupt that exploration. The strategic objective shifts from maintaining a static perimeter to ensuring systemic viability throughout a continuous, algorithmic duel between the formal AI Attacker and the AI Defender.

6. Dynamic Resilience and the Barbell Strategy

Embracing Resilience Theory

Accepting that cybersecurity has transformed into an automated, game-theoretic interaction forces a re-evaluation of human governance. If algorithms are fighting algorithms at machine speed, what is the role of the executive board or the risk manager? When probability can no longer be calculated and scenarios mutate in real time, the traditional corporate practice of residual-risk acceptance becomes an insufficient model. Human governance can no longer focus on preventing all adverse events. Instead, the strategic objective must pivot completely toward resilience theory and the management of deep uncertainty, moving away from narrow financial models toward systems engineering methodologies designed for total risk management (Haimes 1991, Bodeau and Graubart 2016).

This modern alignment shifts the focus from simple structural fortification to continuous functioning and adaptive survival amid structural disruption, a conceptual transition rooted in the ecological resilience paradigms established by Holling (1973). Rather than viewing resilience as merely an engineering metric focused on returning a system to a pre-existing static stasis, modern corporate networks require multi-layered capability frameworks to absorb, adapt, and sustain operational availability despite adverse conditions (Linkov et al. 2013, Linkov et al. 2014). This position aligns with definitions established by

the National Academy of Sciences (2012), which treats disaster and systemic resilience as an explicit institutional imperative. In the classical paradigm, success meant ensuring that an adverse scenario never occurred. In the era of autonomous threats, success means ensuring that when a breach does occur, the organisation can continue to execute its core mission while adapting to the shock. It is the transition to active survivability across hyper-interconnected digital dependencies (Dessavre and Ramirez-Marquez 2015).

Operationalising the Barbell Strategy

To govern an enterprise in this state of constant, unquantifiable exposure, leaders can look to the barbell logic developed by the risk researcher Nassim Nicholas Taleb (2012). While originally designed for financial markets, Taleb's barbell logic provides a highly useful conceptual lens for decision-making under deep, Knightian uncertainty. Applied to cybersecurity governance, this framework suggests combining a highly conservative approach for assets where failure is unacceptable with a highly flexible and adaptive approach at the digital perimeter. A barbell strategy deliberately avoids the middle ground of moderate, calculated risks, which are often based on unstable probability models, creating instead a robust operational posture through two distinct domains: a protected core and an adaptive perimeter.

The conservative side of the barbell requires the absolute safeguarding of the protected core. Governance boards must identify the vital operational assets that the organisation cannot afford to compromise under any circumstance, such as master transactional ledgers, essential customer records, or primary operational control systems. These assets are managed using *strict, unyielding, and highly restrictive validation measures, effectively isolating them from* the fluid, unpredictable outer environment through zero-trust architectures and cryptographic boundaries.

Conversely, the opposite side of the barbell defines the adaptive perimeter. Because the mutations of the threat landscape cannot be predicted, the outer layers of the digital infrastructure must be allowed to dynamically respond and react. This side of the barbell accepts that minor incidents, temporary disruptions, and systemic skirmishes are an inevitable part of operating in an AI-enabled environment. By allowing the automated tools to autonomously isolate compromised segments and adjust tactics in real time at the perimeter, the organisation gains the agility needed to withstand an adaptive adversary. The barbell strategy reconciles human oversight with machine-speed execution. Humans govern the barbell by defining what constitutes the protected core and establishing the operational constraints for the adaptive perimeter, ensuring that even when the rules of the environment are altered by a threat, catastrophic systemic shocks are absorbed.

The preceding chapters describe a transformation not only in the nature of cyber threats and defensive capabilities, but also in the operational logic through which organisations govern them. Figure 2 summarises this transformation by contrasting the classical GRC loop with the continuous feedback loop of autonomous cyber-defence.

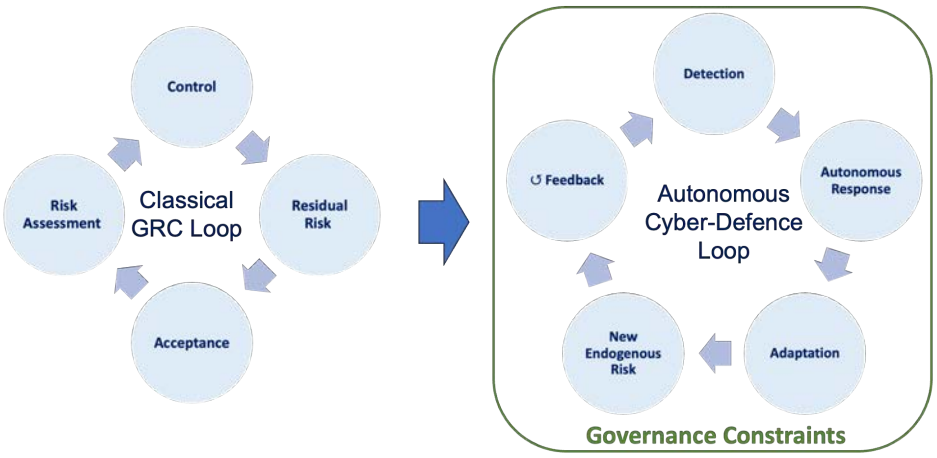

Fig. 2.: Conceptual transformation from the classical GRC loop to the autonomous cyber-defence loop.

The transformation illustrated in Figure 2 can be summarised through five shifts in the underlying logic of GRC. In the classical paradigm, risk assessment provides the basis for selecting controls and evaluating residual exposure. In an autonomous cyber-defence environment, governance increasingly establishes the constraints within which adaptive systems are permitted to operate.

Table 1. *The Transformation from Classical GRC to Autonomous GRC*

Classical GRC	Autonomous GRC
Risk assessment determines controls	Governance manages risk
Controls mitigate assessed risks	Autonomous systems continuously respond
Residual risk is assessed	New endogenous risk continuously emerges
Management accepts residual risk	Management governs autonomy boundaries
Periodic reassessment	Continuous feedback and adaptation

This transformation does not eliminate risk assessment or residual-risk acceptance. Rather, it changes their position within the governance architecture.

Risk assessment becomes an input to defining the conditions and boundaries within which autonomous defence can operate, while management responsibility extends to the behaviour of the defence mechanism itself. The central governance question therefore shifts from determining how much residual risk the organisation is willing to accept to determining within what boundaries autonomous defence can operate safely.

7. The Biological Analogy and the Autoimmune Problem

The Cyber-Immunity Taxonomy

The logical progression from game-theoretic interactions to the barbell strategy brings us to a compelling conceptual bridge, namely the biological immune system. When a digital architecture must defend itself autonomously, adapt to unknown pathogens without human intervention, and maintain core functions during an attack, it ceases to behave like a piece of static engineering, beginning instead to mirror the complex, self-healing mechanisms of a living organism. This alignment provides a distinctive template for modern cyber-defence, moving beyond superficial metaphors to establish a structural correspondence between biological systems and adaptive cyber architectures as conceptualised by Kephart and Chess (2003) in their foundational work on autonomic computing systems.

The human body does not survive in a sterile bubble, nor does it rely on a manual list of known diseases to protect itself. It operates in a state of continuous exposure to mutating viruses and bacteria. The biological immune system achieves resilience by instantly recognising foreign elements, deploying automated defences, and actively repairing damaged tissue, all without conscious human thought. A cyber-system operating under the new paradigm must develop a similar cyber-immunity. To formalise this transition, we must explicitly map the taxonomic equivalents that govern this automated environment, transforming abstract biological functions into precise digital capabilities:

- Pathogen to Cyber Threat: Just as a physical virus acts as a malicious agent invading a host organism, the digital pathogen manifests as a mutating, self-propagating cyber threat or an autonomous adversarial exploit.
- Immune Response to Defensive Response: The immediate mobilisation of cellular defences maps directly to an automated, algorithmic defensive response, triggered at machine speed to contain anomalous network activity.
- Adaptive Immunity to Adaptive Cyber-Defence: Beyond static, hard-coded rules, the system requires an adaptive cyber-defence capable of

altering its tactical posture to neutralise previously unencountered zero-day attack methodologies.

- Immune Memory to Threat and Behavioural Learning: The biological ability to store pathogen signatures for future defence translates into threat and behavioural learning models, ensuring the defensive AI continuously updates its knowledge repository based on past skirmishes.
- Infection to Successful Compromise: Governance must accept that exposure will lead to localised vulnerability breakthroughs. An infection is the exact equivalent of a successful compromise at the digital perimeter.
- Recovery to Cyber Recovery: The final stage of a resilient system is not permanent stasis, but the execution of cyber recovery, allowing the digital infrastructure to programmatically purge the threat and reconstruct baseline configurations without manual intervention.

The Autoimmune Problem and Quarantine Cascades

However, extending this biological taxonomy exposes a profound new governance challenge, specifically the autoimmune problem of autonomous cyber-defence and the systemic collapse of unconstrained shields. In biology, an over-aggressive or miscalibrated immune system can trigger an autoimmune reaction, losing the ability to distinguish between foreign pathogens and the body's own healthy tissue. In an autonomous digital ecosystem, this translates into a highly disruptive maladaptive defensive action, highlighting that when defenders become active system actors, the primary threat of failure frequently transitions from exogenous intrusions to endogenous, uncalibrated control interactions (Young and Leveson 2014).

If an over-reactive defensive AI agent flags standard, high-throughput corporate microservices or legitimate cross-border database synchronisations as active network actions, the result could be a cascade in which successive defensive actions progressively disrupt legitimate business processes, supply-chain connections, or internal services. The system may therefore become unavailable not because an attacker has breached its core, but because its own defence mechanism has incorrectly interpreted legitimate behaviour as hostile. Conversely, an under-calibrated system falls into a state of structural immunodeficiency, leading to a comprehensive defensive failure where the autonomous shield stands idle while the core infrastructure is quietly dismantled by lateral operations.

This reality crystallises the central question of modern risk governance: How can a defence mechanism be sufficiently autonomous to respond to threats at machine speed while remaining sufficiently constrained to prevent it from damaging the system it protects?

Answering this question transforms the conceptual purpose of modern

Governance, Risk, and Compliance. To prevent both structural disruption and absolute vulnerability, the biological organism relies entirely on strict immune regulation. In corporate networks, this regulatory mandate represents the ultimate justification for governance constraints on autonomous defence. Human leadership can no longer operate as a bureaucratic exercise in checking static configuration logs or plotting subjective risk probability heatmaps. Instead, GRC must transform into an architectural engineering framework designed to define the operational limits, the behavioural boundaries, and the absolute cryptographic fail-safes of the cyber-immune system. The primary task of human governance shifts entirely from trying to predict the external attacker to managing the structural regulations of the internal defender, ensuring that the autonomous shield never destroys the corporate organism it was built to protect.

8. Conclusion

The Transformation of Residual Risk

The evolution of the digital landscape from static networks to autonomous, intelligent ecosystems demands an evolution of classical risk management frameworks. The traditional corporate concept of residual-risk acceptance is not going to disappear, nor does the reality of adaptive environments require its elimination. Instead, the strategic baseline changes what residual risk actually means for the enterprise.

Traditional Governance, Risk, and Compliance asks a singular question: how much residual risk an organisation is prepared to accept after technical and operational controls have been applied to mitigate a threat. Autonomous defence introduces a critical, second question that modern boards must confront: what risks are created by the deployment of the controls themselves? Security can no longer be managed as a closed economic calculation or a checkbox compliance exercise focused entirely on exogenous factors. Relying on historical asset distributions, qualitative heat maps, and mathematical probability formulas to stop an Agentic AI attacker is increasingly insufficient. The risk profile shifts from a stationary metric to a dynamic representation of structural dependencies.

The Emergence of Autonomy Boundaries

This profound shift requires modern governance frameworks to introduce a vital conceptual distinction, specifically separating traditional risk acceptance from autonomy acceptance. Classical Governance, Risk, and Compliance has historically focused on a single parameter: how much residual risk are we prepared to accept? However, in an ecosystem driven by algorithmic self-healing,

the corporate board must additionally ask: how much autonomous authority are we prepared to grant to the system managing that risk?

This operational reality demonstrates that a conventional definition of risk appetite is no longer sufficient. To maintain structural control, the board needs to establish explicit autonomy boundaries. These parameters define the exact operational limits of the defensive agent. Rather than allowing an unconstrained algorithm to dictate systemic changes, human oversight must explicitly formalise what the AI can observe, what it can independently decide, what infrastructure segments it can change, and what specific parameters it can isolate or shut down. Crucially, governance must define at what precise structural threshold human intervention becomes mandatory, ensuring that highly consequential defensive actions remain completely reversible. By establishing these boundaries, corporate oversight prevents maladaptive defensive escalations, shifting the role of the auditor from managing abstract external probabilities to validating the structural constraints of internal execution.

The Climax of Modern Governance

To visualise this permanent structural transition, modern governance must evaluate the procedural evolution of the security loop itself. Under classical frameworks, the risk lifecycle is linear and predictable:

Classical GRC:

Threat → Risk → Control → Residual Risk → Acceptance

Under the paradigm of autonomous cyber-defence, the lifecycle shifts into a dynamic, cyclical feedback system:

Autonomous Defence:

Threat → Detection → Autonomous Response → Adaptation → New Endogenous Risk → Governance

As illustrated in Figure 2, the transition from classical GRC to autonomous cyber-defence changes not only the speed and mechanisms of response, but also the object of governance itself.

This does not make risk assessment irrelevant. Rather, it changes its role. In the classical GRC paradigm, risk assessment is the primary mechanism through which organisations determine which controls to deploy and how much re-

sidual risk to accept. In an autonomous cyber-defence environment, however, any assessment represents only a conditional view of a system whose threats, defensive responses and interactions may change continuously. Risk assessment therefore becomes less a mechanism for determining a stable level of acceptable exposure and more a means of establishing structural conditions, critical dependencies and boundaries for autonomous action.

This procedural shift exposes the foundational truth of modern digital oversight: the control is no longer external to the risk; it becomes an adaptive actor within the risk environment. Consequently, risk assessment can no longer remain the sole organising principle of cyber governance. Its role shifts from predicting and accepting a relatively stable exposure to informing the architectural constraints, operational boundaries, and technical fail-safes within which adaptive systems are permitted to operate. Human leadership must therefore govern not only the risks facing the organisation, but also the behaviour of the autonomous mechanisms deployed to manage them. The mandate of the modern director is consequently no longer limited to preventing external intrusion, but extends to regulating these dynamic internal shields, ensuring that the autonomous systems keeping the corporate organism alive never trigger its own fragmentation.

References

Bodeau, D.J. and Graubart, R. 2016. *Cyber Resiliency Metrics: Key Observations*. McLean, VA: The MITRE Corporation.

Cassandras, C.G. and Lafortune, S. 2008. *Introduction to Discrete Event Systems* (2nd Edn.). New York, NY: Springer.

ENISA. 2024. *ENISA Threat Landscape 2024*. Athens: European Union Agency for Cybersecurity.

Haimes, Y.Y. 1991. Total Risk Management. *Risk Analysis* 11 (2): 169–171.

Holling, C.S. 1973. Resilience and Stability of Ecological Systems. *Annual Review of Ecology and Systematics* 4: 1–23.

Kaplan, S. and Garrick, B.J. 1981. On The Quantitative Definition of Risk. *Risk Analysis* 1 (1): 11–27.

Kephart, J.O. and Chess, D.M. 2003. The Vision of Autonomic Computing. *Computer* 36 (1): 41–50.

Kianpour, M., Kowalski, S.J. and Øverby, H. 2021. Systematically Understanding Cybersecurity Economics: A Survey. *Sustainability* 13 (24): 13677.

Knight, F.H. 1921. *Risk, Uncertainty and Profit*. Boston, MA: Hart, Schaffner & Marx; Houghton Mifflin Company.

Linkov, I., Eisenberg, D.A., Plourde, K., Seager, T.P., Allen, J. and Kott, A. 2013. Resilience Metrics for Cyber Systems. *Environment Systems and Decisions* 33 (4): 471–476.

Linkov, I., Bridges, T., Creutzig, F., Decker, J., Fox-Lent, C., Kröger, W., Lambert, J.H., Levermann, A., Montreuil, B., Nathwani, J., Nyer, R., Renn, O., Scharte, B., Scheffler, A., Schreurs, M. and Thiel-Clemen, T. 2014. Changing the Resilience Paradigm. *Nature Climate Change* 4 (6): 407–409.

National Research Council. 2012. *Disaster Resilience: A National Imperative*. Washington, DC: The National Academies Press.

NIST. 2025. *Cybersecurity Framework Profile for Artificial Intelligence (NIST IR 8596 IPD)*. Gaithersburg, MD: National Institute of Standards and Technology.

Taleb, N.N. 2012. *Antifragile: Things That Gain from Disorder*. London: Allen Lane.

Trivedi, K.S. 2002. *Probability and Statistics with Reliability, Queuing, and Computer Science Applications*. New York, NY: John Wiley & Sons.

von Neumann, J. and Morgenstern, O. 1944. *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Young, W. and Leveson, N.G. 2014. An Integrated Approach to Safety and Security Based on Systems Theory. *Communications of the ACM* 57 (2): 31–35.

Avishag Shapira, Simon Shigol, and Asaf Shabtai, 2026, FRAME: Comprehensive Risk Assessment Framework for Adversarial Machine Learning Threats. It is published in Applied Soft Computing; Wiz Research and Irregular (2026).

Dante Gama Dessavre, Jose E. Ramirez-Marquez, Kash Barker, Multidimensional approach to complex system resilience analysis, Reliability Engineering & System Safety, Volume 149, 2016, Pages 34-43, ISSN 0951-8320.

Guo, Y., Liu, P., Ma, W., Deng, Z., Zhu, X., Di, P., Xiao, X., and Wen, S. 2025. Systematic Analysis of MCP Security. arXiv preprint arXiv:2508.12538. https://doi.org/10.48550/arXiv.2508.12538.

Zhao et al. 2025. When MCP Servers Attack: Taxonomy, Feasibility, and Mitigation. arXiv:2509.24272.

Nir Ohfeld, Yuval Avrahami, Wiz Research and Irregular. 2026. Testing AI Agents on Web Security Challenges: What We Learned. 29 January 2026. https://irregular.cv/research/testing-ai-agents-on-web-security-challenges.

CYBERSECURITY FOR ARTIFICIAL INTELLIGENCE

Corrado Aaron Visaggio
University of Foggia, Italy

Abstract
The rapid and pervasive adoption of artificial intelligence systems, particularly large language models, computer vision systems, and autonomous agents, has introduced a new and rapidly evolving attack surface that differs substantially from that of traditional software. As AI systems are increasingly embedded in critical infrastructures, industrial processes, healthcare, finance, and public administration, the consequences of insecure AI extend beyond conventional data breaches to include manipulated decision-making, biased or unsafe outputs, intellectual property theft, and systemic failures with real-world safety implications. There is a growing need for organizations to demonstrate structured risk assessment and mitigation for AI systems, yet the technical and methodological maturity needed to meet these requirements is still developing across both academia and industry.

1. Introduction

Despite growing awareness of AI-specific risks, there remains a significant gap between the pace of AI adoption and the availability of rigorous, validated methods for securing AI systems throughout their lifecycle, from data collection and training to deployment and monitoring. An insecure AI – i.e., an AI-based system that is not properly secured – may cause damage of various kinds: it may introduce discrimination into decisions, allow an attacker to obtain sensitive data belonging to organizations and private citizens, violate the intellectual property of enterprises or governmental confidential information and files, or subvert control of AI-based systems with dramatic consequences: facilitating terrorist attacks, enabling mass surveillance, hindering industrial processes, or taking control of critical infrastructure. This chapter will illustrate the main vulnerabilities and attacks to which AI systems are exposed, along with considerations regarding their feasibility; the impacts of such attacks will then be discussed, together with the techniques and tools at the state of the art that can be used to address the problem.

2. Attacks on Machine Learning

Machine learning is the technology underlying almost the entirety of AI systems. Attacks to machine learning systems can be grouped into five categories:

i) Training set poisoning; ii) Backdoor in the training set; iii) Adversarial example attacks; iv) Model theft; and v) Recovery of sensitive training data.

The usual operation of a machine learning system includes two different phases: the training phase, during which the classifier learns from the data and builds the model of reasoning to use and the testing phase, during which the model is used with data that must be classified.

The first two attacks can be performed during training phase, while the last three attacks can be realized during the test phase.

Training Set Poisoning

An attacker can manipulate data belonging to the training set in order to subvert the predictions of a machine learning model; this is known as a poisoning attack. This type of attack falls into the broader category of supply chain attacks.

It can occur when the attacker is able to access the training dataset that the victim will use to build the model. Two main scenarios enable this attack. In the first scenario, the attacker gains access to the repository where the training dataset is stored. This is achieved through traditional cyberattacks aimed at compromising the victim's perimeter. The countermeasures against this scenario consist of traditional perimeter protection measures.

In the second scenario, the attacker distributes a training dataset that has been built ad hoc to enable the attack. This scenario is particularly effective given the current production approach to machine learning. Indeed, training machine learning systems requires large labelled datasets. Building such datasets is a highly complex and expensive task, so it is common practice to use datasets that are freely available in online repositories. Organizations may lack the time, the data, or the expertise necessary to produce such datasets themselves. Attackers exploit this situation by distributing poisoned datasets.

The effects of this attack can vary: (i) degradation of the machine learning classifier's performance; (ii) induction of a generic error in the model; or (iii) induction of a specific error in the model.

Performance degradation may render a model unusable in practice, while inducing errors compromises the model's effectiveness, as it will produce incorrect predictions or fail to correctly predict specific classes of entities.

Two possible applications of this attack are worth mentioning. A biometric classifier could fail to recognize the face of a device's owner, who would then be denied access to her own device. A facial recognition system could fail to identify the face of a terrorist in an airport lounge. Similarly, a malware detection system might fail to correctly classify malware, allowing attackers to introduce it into an organization's network.

Backdoor in the Training Set

Recently, researchers (Gu et al., 2017) (Ji et al., 2017) have demonstrated that it is possible to create a backdoor and embed it in the training dataset or in a pre-trained model. A backdoor is a hidden function that is added to a system and used to control it, to allow attackers to gain entry, or to make the system perform undesirable or forbidden actions.

A backdoor attack does not alter the model's normal behaviour; however, when a specific trigger is presented to the model, the backdoored instance will be misclassified as the target label desired by the attacker.

The backdoor is injected during the training phase, while during the test phase the attacker submits input containing the trigger. The victim model will then classify the backdoored instance as the attacker's target label.

This attack can be used for different purposes.

The attacker could aim to prevent the recognition of a specific entity. This case is similar to the one mentioned for training set poisoning attacks; the goal is to prevent the correct classification of an entity (a piece of malware, the face of a criminal, a fire in the woods, etc.).

Another scenario involves the misclassification of a positive sample, i.e., the model fails to correctly classify a legitimate entity. Consider an attacker who wants to sabotage a medical device used for classifying CT scans, so that it fails to recognize a specific type of disease. This attack could be used to classify the presence of the disease as its absence.

Adversarial Examples Attacks

Adversarial examples represent noise added to input data, appropriately crafted by an attacker to induce the model to make a mistake when predicting instances.

The term "adversarial example" was introduced by (Szegedy et al., 2014), targeting deep learning algorithms. However, this concept and related methods had been studied much earlier, under the name adversarial machine learning, targeting non-deep machine learning algorithms (Biggio and Roli, 2018), (Dalvi et al., 2004).

These attacks are also referred to as evasion attacks, because they are performed to hinder the detection of malware, spam, network intrusions, and other similar cyber threats.

Adversarial example attacks can be further grouped into two classes (Biggio and Roli, 2018): error-generic attacks, which simply cause the model to produce incorrect classifications; and error-specific attacks, which aim to make the model incorrectly identify the adversarial example as belonging to a specific class.

This kind of attack is particularly useful in the cybersecurity domain, when an attacker wants to conceal the presence of an attack from the classifier.

For the sake of correctness, it is important to note that this kind of attack

is not easy to carry out. First, the attacker is required to have knowledge of the victim's classifier, which is difficult to obtain in practice. Moreover, the success of the attack depends on how much knowledge the attacker possesses.

Several toolboxes for adversarial example generation already exist, such as Cleverhans (Papernot et al., 2018), AdvBox (Goodman et al., 2020), and Adversarial-Playground (Norton et al., 2017), which are mainly intended for research purposes.

It is worth mentioning that adversarial examples can be more effective for some kinds of entities, such as images, and much more difficult to craft for others, such as malware.

In any case, this attack is particularly insidious, even though the technology is not yet mature enough to constitute a short-term risk.

Model Extraction Attacks

Some researchers (Tramèr et al., 2016) showed that an adversary can steal a machine learning model by observing the output labels and confidence levels with respect to chosen inputs. This attack is known as a model extraction attack or model stealing attack.

In the first demonstration of the attack, the model was reconstructed through a series of multiple queries submitted by the user. The interesting feature of this attack is that the threat actor can carry it out simply by submitting normal queries through the model's interface, such as prediction APIs.

Shi et al. (2017) propose a method for stealing a model that builds a functionally equivalent model based on deep learning. It is used in a black-box scenario, where the attacker can only work with the predicted labels from the target model, and employs deep learning to infer and then construct an equivalent model (Shi et al., 2017). Specifically, they use input data to query the target model, and leverage the results produced by the target model to label that input data (Shi et al., 2017). The labelled data is then used to train a model that has the same capabilities as the target model.

However, existing model extraction attacks need to query the target model many times. If the system limits the number of queries, these model extraction attacks may not be able to succeed.

Recovery Sensitive Training Data

There is another pair of attacks that affect the privacy of data in machine learning systems: (i) membership inference attacks, in which the attacker tries to determine whether a specific data sample was used to train the model; and (ii) model inversion attacks, in which the attacker infers information about the training data.

The membership inference attack was introduced by Shokri et al. (2017): the authors demonstrated how an attacker can determine whether a given data

point belongs to the training set of a target model. The membership inference attack proposed in (Shokri et al., 2017) is generic, but the success of membership inference attacks depends on the degree of overfitting of the model (Gu et al., 2017), (Shokri et al., 2017). A well-generalized model can reduce the success of a membership inference attack.

The model inversion attack was first proposed by Fredrikson et al. (2014) in the context of pharmacogenetics tasks. The researchers used black-box access together with additional information about a patient to recover the patient's genomic information. The authors then proposed using the confidence values of predictions to carry out the attack.

As an illustrative example, consider a facial recognition model in which an attacker who knows the user's name can obtain a recognizable image of that user's face..

Defenses

The discussion of protection methods is complex, but the main points can be summarized as follows. In the training phase, the defensive technique against poisoning attacks or backdoor attacks is data sanitization (Koh et al., 2018): the poisoned data is filtered out before being fed into the training phase (Koh et al., 2018).

In the test phase, smoothing of the model's outputs (Papernot et al., 2018) can be adopted as a defense against adversarial examples: this consists of reducing the sensitivity of the model's output to changes in the input. Defense techniques against sensitive information leakage fall into three main categories: distributed learning frameworks, approaches based on traditional cryptographic primitives (e.g., differential privacy, homomorphic encryption), and trusted platform-based approaches.

3. Conclusions

Artificial Intelligence is a technology that is rapidly spreading through our systems, coming to govern an extremely broad spectrum of processes. We will delegate to it the reasoning behind the decisions we make in every domain of human activity: from medicine to scientific research, from the driving of every type of vehicle to military operations, from political programs to artistic production. It is not far-fetched to think that soon each of us will be accompanied by an artificial intelligence assistant supporting us in every daily action.

This means that this technology will manage an enormous amount of data, including sensitive data, will possess vast knowledge, and could become a single point of failure for any process in which it is embedded.

For this reason, it is essential to identify its vulnerabilities – the ways in which it can fail or be exploited by malicious actors to cause harm to citizens, companies, or government agencies.

Today we must focus not only on the development of this technology but, above all, on its governance, and its safe and secure use is a fundamental part of that.

In this chapter, five classes of attacks have been illustrated, but these represent only the results of the scientific community's first studies on this topic.

References

B. Biggio and F. Roli, Wild patterns: Ten years after the rise of adversarial machine learning, Pattern Recognition, vol. 84, Elsevier, 2018.

N. N. Dalvi, P. M. Domingos, Mausam, S. K. Sanghai, and D. Verma, Adversarial classification,' in Proc. 10th ACM SIGKDD Int. Conference of Knowledge Discovery Data Mining, ACM, 2004.

M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing, in Proc. 23rd USENIX Security Symposium, 2014.

T. Gu, B. Dolan-Gavitt, and S. Garg, BadNets: Identifying vulnerabilities in the machine learning model supply chain, 2017, arXiv:1708.06733. [Online]. Available: http://arxiv.org/abs/1708.06733

Y. Ji, X. Zhang, and T. Wang, Backdoor attacks against learning systems, in Proc. of Communications and Network Security (CNS), IEEE, 2017.

P. W. Koh, J. Steinhardt, and P. Liang, Stronger data poisoning attacks break data sanitization defenses, 2018, arXiv:1811.00741. [Online].Available: http://arxiv.org/abs/1811.00741

N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman, A. Kurakin, C. Xie, Y. Sharma, T. Brown, A. Roy, A. Matyasko, V. Behzadan, K. Hambardzumyan, Z. Zhang, Y.-L. Juang, Z. Li, R. Sheatsley, A. Garg, J. Uesato, W. Gierke, Y. Dong, D. Berthelot, P. Hendricks, J. Rauber, R. Long, P. McDaniel, Technical report on the CleverHans v2.1.0 adversarial examples library," arXiv:1610.00768. [Online]. Available: *http://arxiv.org/abs/1610.00768*, 2018.

D. Goodman, X. Hao, Y. Wang, Y. Wu, J. Xiong, and H. Zhang, Advbox: A toolbox to generate adversarial examples that fool neural networks, 2020, arXiv:2001.05574. [Online]. Available: *http://arxiv.org/abs/2001.05574*

A. P. Norton and Y. Qi, Adversarial-playground: A visualization suite showing how adversarial examples fool deep learning, in Proc. IEEE Symposium on Visualization Cyber Security (VizSec), IEEE, 2017.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, Intriguing properties of neural networks, in Proc.2nd International Conference on Learning Representations, OpenReview, 2014.

F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, Stealing machine learning models via prediction APIs, in Proc. 25th USENIX Security Symposium, 2016.

Y. Shi, Y. Sagduyu, and A. Grushin, How to steal a machine learning classifier with deep learning, in Proc. IEEE Int. Symposium on Technology Homeland Security (HST), IEEE, 2017.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov, Membership inference attacks against machine learning models, in Proc. IEEE Symposium on Security and Privacy (SP), IEEE, 2017.

THE FIVE LAYERS OF AI

Mario Beccia
Council of the European Union

Abstract
Generative AI is now embedded in daily routines and in high-risk decisions, from healthcare to military targeting. A large language model is not written but trained: its behaviour is statistical rather than deterministic, its knowledge cannot be inspected line by line and its values are set by whoever trains it. Governance designed for deterministic software does not transfer to these systems, almost all of which are built outside Europe. The chapter proposes a five-layer model that compares the training of a frontier model to the education of a child: foundation, refinement, values, consultation and consumption. The layers belong to different actors (model developer, deployer, application developer), linked by a chain of trust that the downstream actors cannot verify. Each layer carries its own attack surface; the most sophisticated attacks sit at Layers 1 to 3, where auditing is hardest. Read as a map of power, the layers expose European dependencies on foreign hardware, on an offshored refinement workforce and on alignment choices that are jurisdictional and political. The call by leading AI laboratories in September 2026 to pace the frontier and to admit independent evaluators supports the central argument: you cannot govern what you cannot inspect. The consequences are most acute for defence. The AI Act does not apply to military uses, the Alliance's own operational application at Layer 5 is a US commercial product, and protection follows procurement. Sovereignty is therefore the starting condition. Readiness, deterrence and integration among allies are the responses, and they form the pillars of a European defence agenda for AI.

1. A Journey Rooted in the 20th Century

Artificial Intelligence has embedded itself firmly into our daily lives: making an appointment for your car's maintenance, researching potential destinations for your next trip, reviewing your essay for language consistency and improving its bibliography, and identifying targets on a map for a military strike. The use of AI, and in particular generative AI, is by now deeply integrated into our daily routines in areas and through methods that we would never have suspected just a few years ago.

The development of algorithms and much of the theoretical research on AI started a long time ago: Alan Turing's article from 1950, "Computing Machinery and Intelligence", is often referred to as the first piece of academic work on AI. Yet it was only at the end of the last century that usable AI applications started to appear. For instance, the famous chess game between IBM Deep Blue and Garry Kasparov on 10 February 1996 can be considered the first appearance of AI where an application received widespread public attention. Subse-

quently, the development of modern generative AI systems accelerated once the transformer architecture established its technical foundations.

This massive theoretical base remained confined to research labs until advancements in cloud computing provided compute, network and storage capacity as a prerequisite to advance into production-ready applications. Modern large language models (LLMs) are trained on vast datasets such as millions of books, large portions of the public internet, and massive amounts of additional sources. The provisioning of training data would not have been manageable without the large cloud computing resources that are available today.

By the early 2020s, the theoretical foundation and modern computing capabilities allowed for the development of large-scale applications at an unprecedented pace.The result is a set of highly customised applications based on probabilistic (i.e. non-deterministic) mechanisms that cannot be easily traced and analysed and that are almost entirely built outside of Europe.

Nobody today can confidently answer questions such as: Whose AI are you using? Who built it? How was it trained? On whose infrastructure does it run? By whose (moral) values does it decide what it will (and especially, will not) do?

The integration of these AI applications into high-risk decision-making processes is problematic. Their use in defence (targeting, intelligence collection and correlation, options generation), healthcare, critical infrastructure protection, and others raises important questions about how these systems are built, trained, fine-tuned, used, and secured.

AI has quickly become a capability to facilitate strategic autonomy, but policy makers, controllers, lawyers, and managers are unable to absorb this high pace of innovation. The topics of regulation, compliance, technical evaluation, and audit are uniformly lagging as it relates to these applications.

This chapter proposes an analysis model in support of those questions.

2. Statistical Machines, not Software

Traditional software has been a long-running source of sovereignty issues. For example, the Dual-Use Regulation (Regulation EU 2021/821) governs technologies for both civilian and military applications. The software products listed in Annex I of this regulation cannot be exported outside of the EU without prior authorisation. This includes physical (e.g., USB sticks) or intangible (e.g., source code, Internet downloads, etc.) means of transfer. This rationale is based on the regulation's definition of dual-use and the test of whether an application fulfils an established set of criteria. But how can one assess whether the training data of a frontier model is subject to this regulation?

Traditional software is deterministic. For any piece of software that we use,

a human has studied the problem at hand, designed the architecture and developed the code. If a bug is found, it can be fixed by reviewing the analysis as needed and then amending the code. The code base remains unaltered throughout the lifecycle of the application, unless explicit modifications are made using the same process. The very nature of this development process ensures that the same input to the application produces the same output. This approach makes the output of the application predictable, where the application's use, implications, and risks can be understood and predicted.

An LLM is different. The underlying rules are not directly devised by a (human) analyst, they are the result of training the model by exposing it to large amounts of data. The model is taught to predict what comes next through billions of small adjustments to its behaviour. There is no (human) analyst who defined and keeps track of the underlying logic, the sequence of actions needed to answer a question is determined at run time, based on the question itself. The model's answers are thus the result of the interaction between its inner architecture and the data that it was trained on. This is comparable to the way a person's character emerges from the interaction between innate capacity and adaptation to the environment.

The result is that LLMs are non-deterministic, their results are based on probability analysis: there is no determined relation between input and output. The output is generated through probabilistic analysis of the input and its relationship with the training given to the model. An immediate consequence of this structure is that the same input can result in different outputs. The lack of a pre-defined algorithm that establishes a direct relation between input and output makes LLMs highly unpredictable.

Furthermore, there is no source code to inspect for potential bugs or issues, and neither is there a logical rule to point at. The model's knowledge is distributed across billions of parameters in the format of a neural network that is not human-readable. A problem detected in the LLM behaviour cannot be solved with a deterministic action (such as modifying code) but requires a dedicated training session which results in a change of the training parameters. These changes, in turn, increase the complexity of the model and, potentially, augment its non-deterministic and unpredictability traits.

This foundational difference between the deterministic and probabilistic (i.e., statistically generated) approaches is the root of governance difficulties that can be observed with modern AI systems.

Governance Issues

A governance system designed for deterministic software is not directly applicable to statistical algorithms and results in at least two consequences.

The first is the difference between knowledge (used during traditional software development) and reasoning (used by LLMs). Whilst a traditional applica-

tion's "knowledge" directly derives from the analysis performed by the coder(s), a model's "knowledge" is a compressed imprint derived from its training data. It is based on statistical determination and not on a deterministic result from a lookup table. The model can be confidently wrong and cannot confidently explain where its "belief" comes from. The "reasoning" feature, observed in newer frontier models, allows splitting a complex problem into smaller steps. While this improves the explainability of an answer, it is still based on statistical probability and not deterministic precision. For low-stakes tasks, this can be seen as a nuisance with benign consequences. For high-consequence decisions, such as the ones underlying defence processes, it creates a structural hazard: a mechanism that is "usually" right and, occasionally, unpredictably and confidently wrong. A mechanism with these characteristics is difficult to place in processes that demand accountability and where human lives are at stake.

The second consequence is that LLM training is conducted in a largely automated, imperfect, and bias-laden process. Bias is a natural consequence of how the training dataset is collected, consolidated and consumed by the model during its training. Current frontier models are trained predominantly on internet-scale, Western, English-language datasets that inevitably absorb the assumptions, blind spots, and prejudices of the dataset.

An attempt to mitigate the risk generated by this process is to add further training stages, supported by human feedback, and with explicit principles as to what makes the model useful and to constrain what it will do. The result of these stages is in itself imperfect: bias reduction is not a switch that can be turned on or off, but it is a contested, values-rooted, never-finished process that is carried out by humans (or institutions, organisations and companies) who also carry their own biases. Every model encodes a set of choices about what is morally acceptable or not, and someone has provided those choices. The legal and moral scaffolding constructed by these choices becomes an integral element of the model.

These two consequences become even more crucial in defence applications. Frontier models are now used by both parties, attackers who want to conduct a successful operation and defenders who aim to minimise an attack's damage. It is conceivable that a model guides an analyst in deciding which anti-missile system to re-procure after depletion while it is also used by the analyst's opponent to decide which weapon to use for the next attack. It is obvious that the statistical (non-deterministic) process and the embedded values in the LLM highlight moral questions that go well beyond purely academic concerns.

A defence system that utilises a frontier model inherits not only its capability but also its opacity and values, which were defined elsewhere, and which it cannot audit, inspect or predict their impact on the LLM results.

Managing and protecting this complexity from cyber-attacks requires a much deeper understanding of a model's lifecycle.

3. Five Layers

Protecting a traditional software application is a bounded problem. The application has a codebase, data residing in one or more datastores, and it runs on well-established infrastructure perimeters. Modern cybersecurity approaches the protection of the perimeter (and other parts of the infrastructure) with dedicated tools and mechanisms.

The advent of AI applications has shown how insufficient these mechanisms can be. The architecture contains several interconnected parts, a stack of dependencies which have different owners, different failure modes, different performance, and different responsible defenders.

This differentiated architecture can be expressed through five layers of sovereignty for frontier models as depicted in Figure 1.



Fig. 1: Five layers of AI sovereignty

This approach portrays how a frontier model structure can be organised for deeper understanding, how cybersecurity measures needed for each layer can be applied, and how the frontier model lifecycle stages (from base training to inference) can impact other traits.

The frontier model's training can thus be compared to a child's education. A child's mind is shaped in layers, such as the brain he or she is born with, the upbringing, schooling, and languages that the young mind is exposed to. It continues with lived experience and the environment that mould the person's personality and, later, professional experience. The mind is also shaped by parental guidance and societal norms that engrain what is and what is not acceptable. Once the child progresses to becoming an adult, personality and deduction determine the person's actions and define relations with other people.

A frontier model is shaped along the same lines:

- **Layer 1: Foundation.** The architecture and the training dataset constitute the "brain" and the "upbringing" of a person. The model starts to acquire raw capability in a step that requires enormous quantities of data (the size of the dataset) and compute capacity for processing.
- **Layer 2: Refinement.** The model is fine-tuned to become effective at executing tasks and learning to be useful.
- **Layer 3: Values.** Alignment and safety are the main components of this layer. It is comparable to the parental guidance and social norms shaping the personality of a child during teenagerhood. This is where the model is fine-tuned to become *acceptable* rather than merely *capable*.
- **Layer 4: Consultation.** This layer adds specific knowledge to the model and is implemented through techniques, such as Retrieval-Augmented Generation (RAG), that extend its knowledge without altering its initial training. This is comparable to a young adult who, on their first job, uses a manual or similar technical document to complete a single task. Here, personality and education are complemented by additional information without alteration of the underlying information.
- **Layer 5: Consumption.** At this layer, the model is used for a specific application. Much like the human now acting in the world, the user interacts with the model. It can be as simple as a chat window or as complex as a mechanical arm assembling parts in a car manufacturing plant.

Layers 1, 2 and 3 focus on the model's core training. Layer 1 is based on automated techniques that extract knowledge from a large dataset through dedicated algorithms. layers 2 and 3 are mostly based on Reinforcement Learning from Human Feedback (RLHF) , a set of techniques that improve the capability of the model by using direct feedback from human operators.

Layer 4 is mostly based on Retrieval-Augmented Generation (RAG), a technique that overlays additional knowledge such as documents, etc., on top of a trained model while keeping it separate from the model's core training.

The fine-tuning performed at layers 2 and 3 is fundamentally different. At layer 3, the model acquires a set of ethical and moral values, which compare to the ones provided to a child by family, schooling, and the environment. At this layer, the nature of the operator working on the fine-tuning determines most of the objectives and results.

The AI Market

By decomposing AI models through these five layers[1], several hidden angles

[1] From a technical perspective, the five layers constitute a simplification of the structure and technology

in the creation and deployment of an LLM become more visible. For instance, looking at the current AI market structure and the way LLMs are engineered, the responsibility for each layer execution can be described as follows:

- **Layers 1 to 3: model developer.** The first three layers are performed by a model developer who takes care of setting up the vast amount of computing resources needed for the initial training at layer 1 and the refinement done at layer 2. These two layers revolve around solving engineering problems, i.e., obtaining a model that is capable of solving specific problems along stable and (mostly) objective principles. The next layer is still owned by the same model developer, but it is quite different: this is where the model learns to make its actions acceptable and morally sound[2.]
- **Layer 4: deployer.** Layer 4 is owned by a deployer, a private entity, distinct from the creator of the frontier model, that can overlay their own documents or information on top of the LLM and offer a more customised service. The deployer does not directly offer these services to the end user but is able to set up an environment where applications can be developed.
- **Layer 5: end-user application developer.** Layer 5 is owned by an end-user application developer, an entity bridging the gap between the knowledge contained in the model and the needs of the end user.

Some further considerations:

- **The model developer** retains deep technical knowledge of the model. Key elements such as the training dataset used, the training algorithms, the inner functions connecting the neural networks inside the model, are all examples of this knowledge. The model developer has a vested interest in keeping this information secret since it constitutes a competitive advantage over competitors. Even open weight models do not disclose this information, which differentiates open-weight models from open-source models. Nevertheless, this information is crucial for a cybersecurity capability to be able to properly defend against certain types of attacks. It is also crucial to audit and benchmark the success of applied learning reinforced at layers 2 and 3.
- **The model deployer** needs to trust the security measures and proper appli-

behind LLMs; but they constitute an effective tool to assess strategic implications

[2] In a recent essay, Dario Amodei (CEO of Anthropic, one of the most prominent AI model developers) called, among others, for having third-party evaluator access to the training stages of Anthropic's models. In such a scenario, it would be possible for that third party (for instance, a government agency) to audit and evaluate the effectiveness of values alignment performed at layer 3 and, possibly, establish standards and performance tests for this stage.

cation of reinforcement learning performed by the model developer. This trust cannot be based on audits or other traditional methods of inspection.

- **The app developer** needs to trust both the model developer and the deployer, and, again, has no mechanism to audit, verify, or examine the performance of the techniques used in the previous layers.

Cybersecurity Considerations

The attack surface of an AI application can be mapped to the five layers as depicted in Figure 2.



Fig. 2: The attack surface of an LLM across the 5 layers. Each layer represents distinct types of cyber-attacks.

This provides an overview of how cybersecurity for AI departs from cybersecurity for traditional applications. Different classes of attacks apply to each layer and to its responsible defender[3]. Here are some examples:

- **Layer 1: data poisoning.** An adversary implants specific content in the training dataset while evading detection. Due to the large amount of data, the model developer does not have the ability to completely verify the data for accuracy or veracity. Attackers can embed certain behaviours by adding specifically crafted information in the dataset that is often invisible to human checks. This is probably the most difficult attack to detect, as no byte can be inspected, no signature can be checked. It is a given that training

[3] The link between the five layers and cyber attack vectors contained in this chapter constitutes a conceptual simplification, and is based on two AI cybersecurity defensive frameworks, the MITRE ATLAS matrix and the OWASP LLM Top Ten.

data pollution, when performed in a focused and vertical way, is simply undetectable until it is too late.

- **Layer 2: Fine-tuning sabotage.** Human feedback used during reinforcement learning is provided through large numbers of human annotators who are often subcontracted and based offshore. A compromised reinforcement pipeline can embed subtle (biased and undetectable) preferences that survive into deployment.
- **Layer 3: jailbreaks and evasion.** Attacks that trick a model into going beyond its safety training, evading the rules that were instilled during reinforcement learning.
- **Layer 4: prompt and retrieval injection.** An adversary can place malicious content into the dataset used for RAG (trusted by the model's security architecture) and the model uses it as a valid and authoritative source. For instance, altered documents can be maliciously inserted into the RAG dataset, tricking the model into retrieving false information and returning it to the end user.
- **Layer 5: excessive agency and overreliance.** At this level, the model makes use of other tools to accomplish actions requested by the end user, while the attacker has the ability to skip standard security measures (e.g., authentication to a banking API); or, falling for social engineering tricks, the end user authorises the model to take actions that benefit the attacker.

No traditional software architecture has such a complex attack surface where trust is dispersed and hardly verifiable. The value of the five-layer model stems from its ability to keep each layer separate and not collapsed with the other. This offers a more detailed view on each layer's potential issues. Defending an AI application implies defending each layer, owned by different actors, against different threats and different attack vectors.

The most sophisticated attacks do not sit downstream (e.g. prompt injection at layer 4), but upstream at layers 1, 2 and 3 where auditing and verification are hard, if not impossible. Real-life attack techniques tend to merge attacks at different layers into a single, coordinated campaign, making the work of the defenders even harder.

4. The Five Layers and Geopolitics

A further reading key offered by the five layers makes it possible to draw a map of power rather than looking at technical architectures, which is depicted in Figure 3.

LAYER	STRATEGIC DEPENDENCY	WHO CONTROLS IT	EU LEVER
1 FOUNDATION	Infrastructure, Linguistic & cultural hegemony	Model developer architecture (US, China)	Chips Act
2 REFINEMENT	Labor offshoring	Model developer's jurisdiction (US, China)	AI Act
3 VALUES	Whose norms are embedded?	Model developer's legal and ethical environment	AI Act, GDPR, DSA
4 CONSULTATION	Strategic autonomy	Worldwide (e.g. also EU) deployers	Data spaces, GAIA-X
5 CONSUMPTION	User experience capture	App vendors (often, non-EU SaaS)	DSA, procurement, GDPR

Fig. 3: Each layer as a sphere of influence: the strategic dependency, who controls it, and the lever available to Europe.

Each layer offers a view on specific geopolitical dependencies:

- **Layer 1: dataset and hardware producers.** Capability to perform training at layer 1 requires vast amounts of hardware, essentially relying on chip production. Countries such as Taiwan, China and the USA have a strong advantage at this layer, with a gap that is hard to bridge in Europe.
- **Layer 2: specialised workforce.** The refinement of a frontier model is performed by a large, often invisible and often offshored annotation workforce, engaged through subcontractors in jurisdictions outside the model developer's direct control. This is not a minor detail as human reinforcement is where further bias enters a model – whose rules are followed during this stage? Under what guidelines?
- **Layer 3: workforce providing alignment.** Similar concerns pertaining to layer 2 are valid at this level as well and make it even more crucial. At layer 3, the model should align with values and moral judgement, elements that are far from being calibrated at a technical level. The value system present in each society can result in distinct and radically different forms of alignment and cause the model to be (un)able to execute tasks where technical capability was established at previous layers. Bias at layer 3 is essentially jurisdictional and political and deliberately or structurally introduced through the choice of alignment principles and their application by its workforce.

The implication for defence applications is severe. A borrowed frontier model brings foreign values into the core of the command-and-control chain: set at layer 3, by a foreign safety team, refined at layer 2 by a contracted workforce that the (European) defence user cannot see, and trained at layer 1 on data it cannot inspect.

This quickly becomes unacceptable in a command-and-control context, since, for instance, tasks that a system refuses to perform may be as consequential as the ones it executes. Command and control must conform to the Law of Armed

Conflict and national rules of engagement. It must not be defined or implied by a vendor's product where unpredictable behaviour may produce adverse results.

In September 2026, Dario Amodei (Chief Executive Officer of Anthropic; the head of one of the most aggressive laboratories in the field) published a short essay: "We Must Pace the Frontier." He posits that the AI industry (including his own company) should deliberately slow down the rate of AI capability improvements. He does not advocate a freeze but a change of pace to allow for enough restraint to let safety, evaluation, and understanding catch up with capability.

Alongside this argument, Mr. Amodei made a unilateral commitment under which Anthropic grants permanent, employee-level access to independent third-party evaluators. These outside verifiers would be able to confirm whether the company's safety promises are being kept. Mr. Amodei encouraged his competitors to make a similar commitment.

This matches some of the considerations that pertain to layers 2 and 3 above, where the creation of dedicated operators, with the sole purpose of auditing and verifying the validity of the alignment actions performed at those layers, would separate the model developer intent. This improves the performance of Anthropic's models to keep a competitive advantage with his peers while safeguarding society's aims of maintaining an adequate level of safety and trustability of frontier models. Within a day, other AI leaders (Sam Altman of OpenAI and Elon Musk of SpaceXAI) supported or agreed that the frontier should be paced and stated that their companies would also admit external evaluators.

The call for pacing can be seen as support for the argument made in this article, that the complex problems of AI technologies are not rooted in capability issues but rather in the moral and ethical alignment that carries from layers 1 and 2 into the shaped values of layer 3. These problems are unglamorous, slow, and commercially unrewarding, and they are crowded out by a capability race. This may be the reason why the AI industry advocates slowing down the race.

Within the five-layer model, the demand for independent third-party access is a demand for auditability of layers 1 through 3 from outside the model developer's boundary. That is the very demand a European deployer of an AI system should make. The reason is identical to the argument introduced above in this chapter, that you cannot govern what you cannot inspect.

5. The Five Layers: Defence Implications

Three Risks

A pause at the frontier would surely benefit humanity at large; however, it would not resolve Europe's situation. Figure 1.4 depicts three risks that would remain valid:

- At first, restraints applied to AI models are asymmetric: one AI lab

choosing to pace itself does not bind others, nor would it slow down state actors and adversaries who are under no such obligation. The 'gift' of voluntary restraint is advantageous for Europe, but it cannot be assumed or planned for.

- Second, the defensive advantage is first realised by the model's owners: consortia concentrated in US jurisdictions defend their own stacks first, while European entities receive the benefits downstream; as customers rather than peers.
- Third, and most pointedly, protection follows procurement: frontier-derived defensive value reaches users only through applications produced by a small number of providers. Sovereign European LLM applications for Defence (e.g. what Europe should want to build at layer 5) would address this risk, but the EU market is not yet capable of delivering them. The more sovereignty Europe builds at layer 5, the more it carves itself out of a protection it does not control.



Fig. 4: Three risks for Europe's cybersecurity posture implied by frontier models.

An Uncertain Journey

These three risks would become even more pronounced in a scenario where a pacing of development is agreed. Nevertheless, their most serious consequences are for defence, where they converge on four pillars: sovereignty (as the underlying condition) impacting readiness, deterrence and integration among allies. Figure 5 maps these four pillars across the five layers.

LAYER	SOVEREIGNTY *what Europe depends on*	READINESS *what must be held*	DETERRENCE *what must be credible*	INTEGRATION *what must be shared*
1 FOUNDATION *data · compute*	*Chips, compute and frontier training mostly outside the EU*	Compute and energy held as military supplies	Capability hidden in training; hard to signal	Frontier investment and compute pooled among allies
2 REFINEMENT *fine-tuning · RLHF*	*Refinement workforce outside European control*	Versions frozen; every update re-qualified	Credibility rests on refinement no one can verify	Common certification before integration
3 VALUES *alignment · safety*	*Foreign values inside the command chain*	Refusal behavior documented against rules of engagement	Human control at every point of escalation	Joint audits against the Law of Armed Conflict
4 CONSULTATION *RAG · retrieval*	*Sovereign ground: classified corpora stay at home*	Corpus security as force protection	Poison-resistant intelligence supports denial	Protected coalition repositories
5 CONSUMPTION *application · user*	*Foreign-built cockpit: switch-off and alteration risk*	Trained operators; drills with degraded AI	A third-party shield deters only while raised	Alliance-only deployment and a European cockpit

Fig. 5: The five layers and European defense; sovereignty as the condition, readiness, deterrence and integration as the responses.

The AI Act does not apply to AI systems used exclusively for military, defence or national security purposes (Article 2(3)). The main European legal instrument for AI has no impact on the area where the highest-risk use is foreseeable, leaving defence uses to rely on national law, the Law of Armed Conflict and, most importantly, on national procurement regulation.

At the same time, frontier laboratories have become defence suppliers. In June and July 2025, the US Department of Defense awarded OpenAI, Anthropic, Google and xAI contracts of up to $200 million each to develop agentic AI for national security missions. In parallel, Palantir Maven Smart System was deployed as a core NATO capability for command and control, acquired for Allied Command Operations in March 2025. In the terms of the five-layer model, the Alliance's own operational application at layer 5 is a US commercial product.

This directly maps to the third risk identified above: procurement impacts capability, and, in this case, capability relies on non-EU products. The European Commission states this risk plainly in COM(2026) 577: without its own compute, models and data infrastructure capability, Europe remains a vulnerable user of frontier model systems that can be suddenly switched off.

An example from Russia's war of aggression on Ukraine provides factual evidence. In September 2022, a commercial satellite provider declined a Ukrainian request to activate connectivity as far as Sevastopol, a decision its owner later confirmed publicly.

Similarly, a military operation that relies on a model accessed through a foreign provider carries the same risk, with one essential difference: not only can a model be turned off (or denied use), but its behaviour can also be altered or influenced (i.e., at layers 2 or 3) in subtle ways that are hard to detect or audit.

Readiness

The White Paper for European Defence (Readiness 2030, 19 March 2025) lists artificial intelligence, quantum, cyber and electronic warfare among its seven priority capability areas, and the Security Action for Europe instrument (Council Regulation (EU) 2025/1106) provides up to €150 billion in loans, conditional on at least 65% of component costs originating in the Union, the EEA-EFTA states or Ukraine.

The five layers model provides a lens to assess the impact of this programme on AI capabilities. A component-cost threshold has immediate application to hardware, but it does not capture the provenance of a model's training data, the workforce that refined it or the values that aligned it (i.e. does not address issues at layers 1 to 3).

Considering the five layers as an analysis tool, an AI-enabled defence capability should be built upon at least five conditions:

- **Self-determined supply chain**. Compute, energy and model access (e.g. layers 1 and 5) are actual supplies in the same sense as ammunition: they must be stocked, protected and available under contested conditions. For instance, a deployed headquarters that relies on an LLM application reached through a foreign cloud can lose its analytical capabilities when the link is cut, throttled or withdrawn. Operations in disconnected, intermittent or limited-bandwidth environments require LLM applications that run on reliable and sovereign hardware at the edge, based on audited models held under national or allied custody. Open-weight models make this partially achievable; for closed models (such as most frontier models), this is hardly achievable as it depends on the vendor's terms.
- **Configuration control**. An LLM application under full *vendor* control is vulnerable to sudden behavioural changes, for example triggered by an overnight model update altering the refusal behaviour of a targeting-support tool between two phases of the same operation. No commander would accept an equivalent change to a weapon's firmware without proper accreditation. LLM applications for military use should consider frozen versions, refusal behaviour (layers 2 & 3) documented against the rules of engagement, and re-accreditation before any update enters service.
- **Data ownership**. Layer 4 is where sovereign defence information can be protected: classified datasets can be made available through an LLM ap-

plication without leaving national or allied networks, ensuring a higher level of sovereignty. At the same time, layer 4 is exposed to retrieval injection cybersecurity attacks: a single document inserted into a coalition intelligence repository can distort the assessment of every analyst who queries it. The model's cybersecurity becomes an angle of mission security.

- **People**. Operators must be trained on the specific system they use, and specifically on recognising automation bias. Exercises should include phases in which the LLM application is degraded or deliberately compromised, in scenarios where the model becomes a liability and a target.
- **Tempo**. The learning cycle on the Ukrainian battlefield can be estimated at around six weeks, and further research indicates that the length of the cyber tasks which frontier models can complete without human help is doubling over months rather than years[4]. A readiness posture measured in procurement cycles of 5 to 10 years (as in traditional EU defence circles) cannot keep this pace. Advanced cybersecurity techniques such as continuous evaluation and red-teaming should become a readiness function, for instance in the way maintenance is for aircraft. This would greatly accelerate procurement cycles and result in improved readiness levels.

Deterrence

Deterrence rests on *resolve* and *capability* as projected to the adversary and is, thus, only as effective as the projection itself. Defence theory distinguishes deterrence *by punishment*, which threatens unacceptable costs, from deterrence *by denial*, which persuades the adversary that an attack will fail.

AI affects both deterrence techniques, through four effects:

- **Signalling**. A capability that cannot be shown deters no one. Much of an LLM application's advantage lies in layers 1 to 3, where disclosure is hard and erodes the advantage itself: the secrecy that protects a model developer's competitive advantage also limits the ability to signal and, thus, the ability to deter. Credibility also has an internal dimension, since commanders and allies rely on a capability whose refinement (layer 2) cannot be audited or verified. Dependency, by contrast, is quite evident: an adversary can observe which systems a European force runs and who supplies them and can estimate whether a third party would keep them running in a crisis or conflict.
- **Attribution**. Deterrence by punishment requires identifying the attacker with enough confidence to justify a response. In cyberspace, attribution was already difficult before the introduction of AI. With modern

[4] https://www.aisi.gov.uk/blog/how-fast-is-autonomous-ai-cyber-capability-advancing

LLM applications, AI-generated code, synthetic personas and automated campaigns make it even harder. Since its 2014 Wales Summit, NATO has held that a cyber-attack can lead to the invocation of Article 5; Article 42(Snyder, 1961) of the Treaty on European Union contains a mutual assistance clause. Both depend on an attribution that AI makes more complex, slower and less certain.

- **Escalation.** The speed implied by the use of AI by both defender and attacker compresses decision time and significantly raises the risk of escalation through miscalculation. In a scenario where automated systems respond to each other faster than commanders can intervene, this risk is inherently higher. The United States and China affirmed, in November 2024, the need to maintain human control over the decision to use nuclear weapons (Reuters, 2024). The same principle (a human decision at every point of escalation, enforced at layer 3 and visible at layer 5) has been proposed for extension to conventional and cyber operations (Reuters , 2026).
- **Denial**. With punishment weakened by attribution and escalation risks, a more credible deterrence path is deterrence by denial. Capabilities such as resilient networks, vulnerabilities patched at machine speed, defence datasets that resist poisoning and LLM application defences that make an attack unlikely to succeed increase the cost for the adversary and deny benefits of attacks. Denial, though, requires *sovereign control* over the defender's infrastructure: a capability partially operated by a third party outside the control of the defender deters only as long as that party can credibly declare its alignment with the defender's objectives.

Integration Among Allies

No EU Member State can fund frontier capabilities or LLM applications alone: the European Commission estimates the investment needed in the hundreds of billions of euro. At the Ankara Summit (7–8 July 2026), NATO Allies called for wider adoption of AI-enabled capabilities and for an interoperable Transatlantic Warfighting Cloud.

Interoperability between allies, however, reproduces the trust chain described above: an Ally deploying another Ally's model is comparable to a deployer trusting a developer it cannot audit. Integration must therefore extend beyond joint procurement to *shared inspection*: pooled investment in frontier models and compute capacity, building on shared and trusted efforts. Examples are the Commission's Frontier AI Grand Challenge for an open-source model in all 24 EU languages; and the proposal for a common certification scheme for AI components before they enter command-and-control systems, for which the certification framework of the Cybersecurity Act (Regulation (EU) 2019/881) offers a model of mutual recognition.

Additional efforts could include joint audits of layer 3 refusal behaviour against the Law of Armed Conflict and national rules of engagement; protected coalition repositories at layer 4; and deployment confined to the Alliance, with no inference routes outside it and a European set of applications at layer 5.

Initiatives such as the working group announced by the Commission with ENISA and the European Defence Agency, and the forthcoming NATO Centre of Excellence on Artificial Intelligence, are paving the way.

6. The Way Ahead

Sovereignty, readiness, deterrence and integration among Allies constitute the foundational pillars of a strategic agenda that, under the lens of the five layers model, would provide an effective defence advantage.

European governments should use any means (including the pause afforded by a paced frontier development) to build their next generation defence tools along those foundational pillars and create sovereign capabilities across the five layers, including auditable values in high-risk settings, and a layer 5 surface for the humans who actually take decisions.

A complex endeavour, but the only way to ensure a sovereign future for the Union.

References

Altman, S. 2024. The Intelligence Age. Available online at personal blog (Last accessed: October 11, 2026).
Altman, S. 2025a. Reflections. Available online at personal blog (Last accessed: October 11, 2026).
Altman, S. 2025b. Three Observations. Available online at personal blog (Last accessed: October 11, 2026).
Altman, S. 2025c. The Gentle Singularity. Available online at personal blog (Last accessed: October 11, 2026).
Amodei, D. 2024. Machines of Loving Grace. Available online at personal blog (Last accessed: October 11, 2026).
Amodei, D. 2025. The Urgency of Interpretability. Available online at personal blog (Last accessed: October 11, 2026).
Amodei, D. 2026a. The Adolescence of Technology. Available online at personal blog (Last accessed: October 11, 2026).
Amodei, D. 2026b. "We Must Pace the Frontier". Personal essay published on 12 September 2026, reported in Axios.
Bremmer, I. and Suleyman, M. 2023. Much Ado About Management. The AI Power Paradox. Foreign Affairs (September/October issue).
Christiano, P., Leike, J., Brown, T.B., Martic, M., Legg, S. and Amodei, D. 2017. Deep RL from Human Preferences. NeurIPS Proceedings.
Council Regulation (EU) 2025/1106. 2025. Establishing the Security Action for Europe (SAFE) instrument (27 May 2025).
Draghi, M. 2024. The Future of European Competitiveness. Report presented to the European Commission (September 2024).
European Commission. 2026a. Action Plan on Cybersecurity and Artificial Intelligence. COM(2026) 577 final (7 July 2026).
European Commission. 2026b. Announcement of the Frontier AI Grand Challenge winner (EUROPA consortium). Available online at official portal (Last accessed: October 11, 2026).
European Commission and High Representative. 2025. White Paper for European Defence – Readiness 2030 (19 March 2025).
Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. Minds and Machines 30 (1): 20–35.
Isaacson, W. 2023. Elon Musk. New York: Simon & Schuster.
Lewis, P., Perez, E., Piktus, A., Petroni, F., Lewis, M., Riedel, S. and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS Proceedings.
Mensch, A. 2026. Testimony to the French National Assembly inquiry commission on structural digital dependencies (12 May 2026).
Mistral AI. 2026. European AI: A playbook to own it. White paper, published April 2026.
NATO. 2014. Wales Summit Declaration. Issued by the Heads of State and Government participating in the meeting of the North Atlantic Council in Wales (5 September 2014).

NATO. 2026. Ankara Summit Declaration. Issued by the Heads of State and Government participating in the meeting of the North Atlantic Council in Ankara (8 July 2026).
NATO Communications and Information Agency. 2025. "NATO acquires AI-enabled Warfighting System" (April 2025).
Ouyang, L., Lowe, J., Williams, T. and OpenAI Safety Team. 2022. Training Language Models to Follow Instructions with Human Feedback. InstructGPT / OpenAI Research (March 2022).
Regulation (EU) 2019/881. 2019. The Cybersecurity Act. Official Journal of the European Union.
Regulation (EU) 2021/821. 2021. Setting up a Union regime for the control of exports, brokering, technical assistance, transit and transfer of dual-use items. Official Journal of the European Union.
Regulation (EU) 2024/1689. 2024. The Artificial Intelligence Act (AI Act), Article 2(3). Official Journal of the European Union.
RUSI. 2024. "UK firm supports Ukrainian armed forces in drone tech race". In the News, citing Reuters.
Snyder, G.H. 2011. Deterrence and Defense: Toward a Theory of National Security. Princeton, NJ: Princeton University Press.
Suleyman, M. 2023. The Coming Wave. New York: Crown Publishing.
The White House. 2024. Readout of the meeting between President Biden and President Xi in Lima (16 November 2024).
Treaty on European Union. 1992. Article 42(7) [Mutual Assistance Clause]. Official Journal of the European Union.
Turing, A.M. 1950. Computing Machinery and Intelligence. Mind LIX (236): 433–460.
US Department of Defense. 2025. Chief Digital and Artificial Intelligence Office, frontier AI contract awards; reported in DefenseScoop (14 July 2025).
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. 2017. Attention Is All You Need. NeurIPS Proceedings 30 (1): 5998–6008.

DISCLAIMER:
This text was drafted in a personal capacity. The views and opinions of the author expressed herein should therefore not be interpreted as stating or reflecting those of his employers or any organization of which he is a member.

CYBERSECURITY FOR CONNECTED MEDICAL CARE: LESSONS FROM THE CYMEDSEC RESEARCH PROJECT ON INTERNET OF MEDICAL THINGS

Cristine Skropke
Secunet Security Networks AG, Germany

Florian Sonnenschein
Secunet Security Networks AG, Germany

Torsten Redlich
Secunet Security Networks AG, Germany

Abstract
Connected medical devices can improve care in hospitals and at home, but they also create dependencies across equipment, digital services and organisations. A device may still measure correctly while missing, delayed or altered information puts the care pathway at risk. The practical challenge is to protect the complete service and establish who is responsible for maintaining that protection, particularly as care moves beyond managed hospital environments. This chapter draws on the CYMEDSEC project to explore an approach built around a gateway that manages connections near a medical device. As a separately maintained security layer, it can reduce exposure and help protect equipment that remains clinically useful after its communication software becomes difficult to update. The project combined threat modelling, architecture development and practical testing to investigate how separating medical functions, additional services and security functions can limit the consequences of an attack. Security reviews and penetration tests supported key design decisions while revealing weaknesses in authentication, credentials and administrative access in the tested configurations. These findings highlight why safeguards must be evaluated as configured and operated, with clear processes for updates, access management and recovery. They also show why established security technologies need to be translated into solutions that organisations can deploy and maintain reliably. The chapter identifies three priorities for wider adoption: common European security baselines that guide development and procurement; long-term maintenance offered as a defined service with clear support commitments; and repeatable testing under realistic operating conditions. For manufacturers, healthcare providers and service operators, the central lesson is to agree before deployment who owns each protection, what evidence demonstrates that it works and how it will be maintained. CYMEDSEC illustrates how technical boundaries, clinical needs and operational responsibilities can be addressed together to support resilient connected care.

1. Innovation in Medical Technology: from Individual Devices to Connected Care

Operating theatres, intensive care units and diagnostic departments are now equipped with highly advanced devices, from real-time imaging to adaptive ventilators and continuous patient monitoring. Yet much of this innovation remains confined to individual devices. Many systems still operate in isolation, and information often travels by hand, on paper or through the attention of clinical staff. This will change. The next stage of innovation lies in connecting devices with one another and with clinical processes: from networked operating theatres to entire process chains extending from admission to aftercare at home. Staff shortages, growing complexity, cost pressure and medical expectations make this development necessary, and it is a welcome one. But connecting medical systems, which thereby become Internet of Medical Things (IoMT) devices, also connects their risks. Networked care brings together security domains that were traditionally considered separately: the functional safety of devices, the cybersecurity of networks and software, and the protection of sensitive health data. Where these domains and the organisations responsible for them meet, a cyber attack can become a patient safety incident. The benefits of connected care will only be realised if these risks are addressed from the outset.

What the ENISA Threat Landscape 2026 Tells us About Connected Care

The ENISA Threat Landscape 2026, published in September 2026, underlines why security for connected care cannot be postponed. ENISA analysed 8.257 incidents recorded between 1 January and 31 December 2025, drawn mainly from open sources and anonymised information shared by Member States and partners. In incident reports submitted by Member States under NIS2, health ranked among the five most affected sectors, alongside public administration, digital infrastructure, banking and transport. Ransomware continued to disrupt the delivery of care directly, as several hospital cases illustrate.

Three findings are particularly relevant for the Internet of Medical Things. First, where an intrusion vector could be identified in unauthorised access incidents, more than 60 percent involved exploited vulnerabilities, with misconfiguration and accidental exposure accounting for about a fifth. Second, attackers repeatedly compromised supply chains, third-party providers and cloud environments, causing large-scale incidents. Third, ransomware operators exploited weaknesses in VPN and firewall appliances. Network-facing devices at the edge of care infrastructure are therefore prime targets. Vulnerability management, secure updates and clear supplier responsibilities are not optional features; they are direct answers to the threats Europe is observing today.

2. Why Security Depends on the Whole Care System

Connected medical care depends on devices, networks, remote services and the organisations responsible for them. In the CYMEDSEC project, we investigated how an edge gateway, a computer that manages connections near a medical device, can help protect this wider system. We developed a threat model, used it to guide the proposed architecture and tested the resulting concepts. Our research shows why technical boundaries, maintenance responsibilities and practical validation belong together.

This question is becoming more pressing as care moves beyond the hospital. Home care and hospital-to-home models allow patients to be discharged earlier and monitored remotely, and they are increasingly seen as a way to relieve strained hospital capacity. But when a medical device leaves the clinical environment, it also leaves the protections that environment provides. In a hospital, devices operate in managed networks, under the supervision of IT and medical technology staff. At home, the same device may share a consumer router with laptops, smart TVs and games consoles, depend on a residential internet connection, and be handled by patients or relatives with no technical training. No one on site is responsible for updates, configuration or noticing that something has gone wrong. At the same time, responsibility becomes more fragmented. The device manufacturer, the discharging hospital, a home care provider, a telemonitoring service and possibly a network operator each control part of the chain, yet none of them controls the environment in which the device actually runs. This makes it all the more important that everyone fulfills their security responsibilities.

3. When Missing Data Becomes a Clinical Risk

Consider a patient whose condition is monitored at home. A medical device collects measurements and sends them through a gateway to a service used by a clinical team. The device, gateway and remote service may come from different suppliers. The patient relies on them working together.

This illustrative example belongs to the Internet of Medical Things, or IoMT: medical devices connected to other systems to support care. Similar arrangements exist in hospitals, where equipment exchanges information with local applications and external services. These systems are distributed across locations and organisations. They become critical when care depends on receiving correct information at the right time.

Now imagine that a measurement does not arrive. The device might have lost its connection. An update might have failed, or an attacker might have in-

terrupted communication. Staff need to distinguish missing information from a reassuring result and know when to contact the patient. A device that still measures correctly does not, by itself, ensure that the care pathway works.

Other risks are less visible. Someone could read private health information, alter a measurement or use a stolen administrator account to affect many installations. Security must therefore protect information against disclosure and alteration while keeping the service available.

Replacing older equipment is not always a practical answer. A device may remain clinically useful after its communication software becomes difficult to maintain. A gateway offers a way to control its network connections through a separately maintained system. This can reduce exposure, although the device and its local connection still need assessment.

The starting point is the clinical dependency. Before choosing security technology, establish what missing, delayed or incorrect information would mean for the people using the service.

Where delayed delivery is clinically acceptable, a gateway could store measurements during a short outage and send them later. For time-critical monitoring, that may be insufficient. Teams must agree how staff are alerted and when care switches to a fallback procedure.

4. Towards a Shared European Understanding of Cybersecurity

Medical technology is a European product in the fullest sense. A patient monitor may be designed in Germany, run software developed in Finland, be manufactured with components from several member states, connect to a cloud service hosted in Ireland and be used in a hospital in Italy. This heterogeneity becomes more critical as care becomes more connected. Security is no longer determined by a single device, but by the interaction of manufacturers, software vendors, infrastructure and service providers, operators and clinical staff, each controlling part of the chain and often subject to different national regimes. Where expectations differ, gaps appear at the interfaces between these actors, and it is there that attacks are most likely to succeed. A common European understanding of adequate cybersecurity, and a common basis for deciding whether a product may be used, is therefore a practical precondition for connected care. The next step is alignment: requirements that allow evidence produced once to satisfy several obligations, a common model for dividing responsibilities between actors, and mutual recognition of assessments across member states. The goal is not to replace national responsibility for healthcare, but to ensure that a product proven secure in one member state does not have to prove itself again in every other.

5. The Essential European Requirements

Europe has already laid important foundations. ENISA, established on a permanent basis by the Cybersecurity Act, is building European certification schemes; NIS2 has harmonised obligations for healthcare providers and device manufacturers; and international standards such as IEC 81001-5-1, IEC 80001-1 and IEC 62443 offer a shared technical language. European research programmes and initiatives help to translate these frameworks into practice. Yet different instruments still apply to different parts of the same care system, with differing terminology, timelines and conformity procedures.

For connected healthcare, three responsibilities are especially important: keeping medical devices safe, protecting health information and maintaining dependable services.

The Medical Devices Regulation (MDR) and In Vitro Diagnostic Medical Devices Regulation (IVDR) address device safety and performance, including software security and risk management. Manufacturers must consider the environment in which their devices operate. Security therefore needs attention throughout development, deployment and maintenance.

The General Data Protection Regulation (GDPR) requires lawful handling of personal data and appropriate protection. Teams must establish who needs access to measurements, why the information is collected and how long it is retained. Encrypting a connection is only part of that work.

The NIS2 Directive addresses organisational resilience, including cybersecurity risk management, incident reporting and supply chain security for healthcare entities within scope. Its applicability depends on the organisation and national implementation.

The Cyber Resilience Act (CRA) addresses the security of products with digital elements, including secure design, vulnerability handling and security updates during the support period. Products covered by the MDR or IVDR are excluded from its scope. General-purpose gateways and software used in healthcare may nevertheless fall within it and need separate assessment. For these products, technical protections must be supported by evidence and processes for maintaining security after deployment.

For connected care, the planned European Cybersecurity Certification Scheme for Cloud Services (EUCS) is particularly relevant, as it would provide a common, EU-wide basis for assessing the cloud and remote services on which telemonitoring, remote maintenance and hospital-to-home models increasingly depend.

The practical implication is to connect engineering decisions with operational responsibility. A software update needs technical checks, an agreed deployment process and a recovery plan. A gateway can support these tasks, but the complete service still needs assessment against its applicable obligations.

CYMEDSEC: securing the Internet of Medical Things from threat model to tested architecture

The CYMEDSEC research project brings together medical technology, cybersecurity and regulatory research to investigate how connected medical devices, the Internet of Medical Things, can be protected in real care settings. Our scenarios range from remote patient monitoring to critical care. At the centre of our work is an edge gateway operating close to the medical device. We examined how additional communication services can run alongside a medical device while their access to the security platform remains limited. The CYMEDSEC research followed three connected activities: threat modelling, architecture development and concept validation.

6. Threat Modelling: Knowing what to Protect

We began with a threat model, a structured description of what needs protection, what could go wrong and which safeguards could address it. Starting from a generic gateway profile, we analysed its components, interfaces and data flows in healthcare settings. Using the STRIDE method, we systematically considered false identities, data manipulation, disputed actions, exposed information, interrupted services and excessive access rights, including threats to remote management and to communication over untrusted networks.

Each identified threat was translated into a security requirement. The risk of an application gaining too much control, for example, called for strict separation between workloads and security functions; risks to transmitted data called for protected communication. The threat model thus gave every architectural protection a documented rationale and a basis for evaluating the design.

7. From Research Findings to Standards, Offerings and Practical Implementation

Our research in CYMEDSEC confirmed that the building blocks for securing connected medical care exist. Secure boot, hardware-bound keys, mutual authentication, signed updates, isolation and network segmentation are established technologies. The challenge lies elsewhere: in most care settings, these measures are neither consistently required, nor available as ready-to-use offerings, nor simple enough to be operated reliably by organisations without dedicated security teams. If connected care is to scale, the insights from research must be translated into three things: common standards that define what adequate protection means, offerings that make this protection available as a service, and implementations that are easy to deploy and maintain.

8. Security Features need a Common Reference

We derived six security features for an edge gateway: a hardware root of trust with encrypted persistent data, cryptographic device identity, resilient updates, defence in depth, software bills of materials with vulnerability management, and network policy enforcement. Each of these is well understood on its own. What is missing is a shared reference that describes them as a baseline for connected medical infrastructure, so that manufacturers, operators and procurers can refer to the same set of requirements. Existing standards such as IEC 81001-5-1 for health software and IEC 62443 for connected systems provide a solid foundation. They should be applied consistently to the gateways and platforms that connect medical devices, not only to the devices themselves, and reflected in procurement criteria. A baseline of this kind would also make it easier to answer the decisive question for each feature: who owns it, what evidence shows that it works, and how it will be maintained.

Long-Term Maintenance Must Become a Defined Service

A system installed in a patient's home may need updates for years after deployment. Our work showed that signed updates and redundant system images address two distinct risks: accepting an unauthorised package and losing a working installation. Yet these technical mechanisms are only effective within a functioning process of release checks, protected signing keys and controlled recovery. Most healthcare providers cannot build and operate such processes themselves. Long-term maintenance should therefore be offered as a defined service with clear commitments on support periods, update frequency and recovery procedures. Contractually fixed support periods, rather than technical possibilities alone, give operators the planning certainty they need.

Validation Must Be Repeatable, not a one-off Exercise

Our security reviews and penetration tests confirmed key design decisions, but they also revealed weaknesses in authentication, credentials and administrative access in the tested configurations. The lesson is that protections must be tested as configured and operated, not merely as present. For practice, this means that validation should not remain a one-off research activity. Standardised test profiles, reference configurations with secure defaults and accessible testing and certification infrastructure would allow manufacturers and operators to verify their systems regularly and comparably. Particular attention belongs to central management: it makes large fleets manageable, but concentrates authority. Requirements for management accounts, access routes and activity records should be part of every standard and every offering.

9. Conclusion

The Project Team around CYMEDSEC shows that securing the Internet of Medical Things requires more than individual protective mechanisms. A systematic threat model provides the rationale for each protection; an architecture that separates medical functions, additional services and security functions limits the impact of compromise; and realistic testing reveals whether protections work as configured and at acceptable cost. Above all, security depends on clear responsibilities between supplier and operator. These should be agreed before installations are deployed and become difficult to reach.

Acknowledgement

Our work in the CYMEDSEC project is funded by the European Union under grant agreement No 101094218. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

References

European Union Agency for Cybersecurity (ENISA). (2026). *ENISA Threat Landscape 2026*.
https://www.enisa.europa.eu/publications/enisa-threat-landscape-2026
CYMEDSEC Consortium. (n.d.). *CYMEDSEC project*. Accessed 13 October 2026.
https://cymedsec.eu/
International Electrotechnical Commission (IEC). (2021). *IEC 81001-5-1:2021. Health software and health IT systems safety, effectiveness and security – Part 5-1: Security – Activities in the product life cycle* (1st ed.).
https://webstore.iec.ch/en/publication/63293
International Electrotechnical Commission (IEC). (2021). *IEC 80001-1:2021. Application of risk management for IT-networks incorporating medical devices – Part 1: Safety, effectiveness and security in the implementation and use of connected medical devices or connected health software* (2nd ed.).
https://webstore.iec.ch/en/publication/34263
International Society of Automation (ISA). (n.d.). *ISA/IEC 62443 series of standards*. Accessed 13 October 2026.
https://www.isa.org/standards-and-publications/isa-standards/isa-iec-62443-series-of-standards
European Parliament and Council of the European Union. (2017). *Regulation (EU) 2017/746 of 5 April 2017 on in vitro diagnostic medical devices (IVDR)*. Official Journal of the European Union, L 117, 5 May 2017.
https://eur-lex.europa.eu/eli/reg/2017/746/oj/eng
European Parliament and Council of the European Union. (2017). *Regulation (EU) 2017/745 of 5 April 2017 on medical devices (MDR)*. Official Journal of the European Union, L 117, 5 May 2017.
https://eur-lex.europa.eu/eli/reg/2017/745/oj/eng
European Parliament and Council of the European Union. (2022). *Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS 2 Directive)*. Official Journal of the European Union, L 333, 27 December 2022.
https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng
Microsoft. (2022). *Microsoft Threat Modeling Tool threats*. Section "STRIDE model". Microsoft Learn. Accessed 13 October 2026.
https://learn.microsoft.com/en-us/azure/security/develop/threat-modeling-tool-threats
European Parliament and Council of the European Union. (2024). *Regulation (EU) 2024/2847 of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act)*. Official Journal of the European Union, L 2024/2847, 20 November 2024.
https://eur-lex.europa.eu/eli/reg/2024/2847/oj/eng
European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 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 (General

Data Protection Regulation). Official Journal of the European Union, L 119, 4 May 2016. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng

European Parliament and Council of the European Union. (2019). Regulation (EU) 2019/881 of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification (Cybersecurity Act). Official Journal of the European Union, L 151, 7 June 2019. https://eur-lex.europa.eu/eli/reg/2019/881/oj/eng

CYBER-SOCIAL SECURITY

Vita Santa Barletta
Department of Computer Science, University of Bari Aldo Moro

Danilo Caivano
Department of Computer Science, University of Bari Aldo Moro

Abstract
The rapid convergence between cyberspace and society is reshaping the nature of cyber threats, which increasingly transcend technical infrastructures to influence human behavior, public perception, organizational processes, and critical systems. To address this evolution, Cyber-Social Security (CSS) has emerged as a multidisciplinary paradigm capable of integrating cybersecurity, artificial intelligence, cognitive and social sciences, linguistics, ethics, and legal studies into a unified framework. Unlike traditional technology-centric models, CSS places people, social interactions, and societal trust at the center of the security discourse, analyzing risks across four interrelated domains: cyber, informational, cognitive, and physical. This paradigm shift exposes significant limitations in current Information Security Risk Management methodologies, such as the Common Vulnerability Scoring System (CVSS), which fail to quantify cognitive manipulation or cascading social impacts. To address this gap, this research advocates the necessity of redefining cyber-risk scoring by extending existing standards through cyber-social metrics capable of quantifying both latent socio-technical vulnerabilities and realized cross-domain impacts.

1. Introduction

The progressive convergence of digital infrastructures, artificial intelligence, online platforms, information ecosystems and societal processes is challenging the traditional boundaries of cybersecurity. Digital technologies no longer constitute a separate environment surrounding social activity; rather, they increasingly mediate communication, mobility, access to essential services, public administration, organizational decision-making and interpersonal relationships (Costa et al., 2025). As a consequence, cyber incidents can no longer be understood exclusively as events affecting hardware, software, networks or data.

A technical compromise may represent only the initial stage of a much broader trajectory. Compromised information can be manipulated or selectively disclosed; manipulated content may influence the perceptions of individuals or communities; altered perceptions may affect decision-making; and decisions based on distorted information may eventually generate economic, organizational or physical consequences (Zegarow and Bartuzi, 2024). Research on cy-

ber harms has already shown that the effects of cyberattacks can propagate beyond their initial technical manifestation and generate cascading organizational and societal consequences (Agrafiotis et al., 2018). At the same time, automated influence, generative artificial intelligence and synthetic media increasingly allow adversarial actors to operate directly on information environments and cognitive processes without necessarily exploiting a conventional software vulnerability (Rajtmajer and Susser, 2020).

This transformation changes not only the threat landscape but also the object of cybersecurity governance. The relevant question is no longer simply whether an information system remains confidential, integral and available. Security increasingly concerns whether the information on which decisions are based is authentic, whether individuals and organizations retain sufficient decision-making autonomy, whether trust relationships can be exploited, and whether digital events can propagate into disruption of societal functions (Austin and Withers, 2020).

Cyber-Social Security (CSS) provides an analytical perspective for addressing this transformation. CSS conceptualizes cyberspace as part of a broader socio-digital ecosystem and examines security through four interconnected domains: *Cyber*, *Informational*, *Cognitive* and *Physical*. Rather than treating these dimensions as autonomous areas of security, CSS focuses on their interdependencies and on the mechanisms through which effects can move from one domain to another (Antoniol et al., 2025). This perspective is consistent with the multidisciplinary foundation of CSS, which combines cybersecurity with human factors, artificial intelligence, social and cognitive sciences, economics, law and other disciplines and reinterprets traditional Detection, Response and Prevention functions in cyber-social contexts (Barletta et al., 2024a).

The relevance of this approach extends beyond cybersecurity engineering. Once the protected environment comprises individuals, information flows, institutional trust, decision-making and essential services, cybersecurity becomes inherently connected to governance and law. This is particularly evident within the European Union, where cybersecurity, artificial intelligence, online platforms, critical infrastructure and product security are governed by different but increasingly interconnected regulatory instruments.

The NIS2 Directive (European Parliament and Council, 2022a) establishes risk-management and incident-management requirements for essential and important entities and explicitly adopts an all-hazards approach encompassing technical, operational and organizational measures. The Critical Entities Resilience Directive (CER) broadens resilience beyond cyber risk and explicitly considers cross-sectoral dependencies, hybrid threats and coordination between cyber and non-cyber competent authorities (European Parliament and Council, 2022b). The Digital Services Act (DSA) introduces systemic-risk obligations

that extend into fundamental rights, civic discourse, electoral processes, public security and physical and mental well-being (European Parliament and Council, 2022c). The Artificial Intelligence Act addresses manipulative AI practices and transparency for synthetic content and deepfakes (European Parliament and Council, 2024a).

These instruments address different legal objects and impose obligations on different actors; they should not be conflated. Nevertheless, taken together, they reveal an evolution towards a more systemic understanding of digital risk.

Therefore, this paper argues that Cyber-Social Security can provide a conceptual bridge between existing cybersecurity metrics and the European regulatory ecosystem. Rather than focusing exclusively on technical severity, CSS extends risk assessment to include latent socio-technical susceptibility, cross-domain propagation, and realized societal impact. This governance-oriented perspective enables dependencies among *Technical Vulnerabilities*, *Information Integrity*, *Decision Autonomy*, *Systemic Resilience*, and physical consequences to be assessed within a common analytical framework.

The contribution is therefore threefold. First, it formalizes the transition from a techno-centric interpretation of cybersecurity towards a *multi-domain cyber-social perspective*. Second, it introduces the concepts of *Influence Surface* and *Propagation Surface* to complement the traditional *Attack Surface* and to capture how cyber-mediated effects propagate through socio-technical environments. Third, it investigates how traditional cybersecurity metrics can be conceptually extended beyond technical severity to incorporate *cyber-social dimensions of risk*, including latent socio-technical susceptibility, cross-domain propagation, and realized societal impact, thereby providing a governance-oriented basis for connecting technical security, information integrity, decision autonomy, systemic resilience, and European regulatory obligations.

2. Methodological Approach

This study adopts an interdisciplinary conceptual approach combining socio-technical cybersecurity analysis with a functional examination of the European digital regulatory ecosystem. The goal is not to provide a doctrinal interpretation of individual legal provisions or to establish a direct equivalence between Cyber-Social Security domains and specific regulatory instruments. Rather, the methodology investigates how cyber-social threats can propagate across technical, informational, cognitive, organizational and physical dimensions, and how these dimensions intersect with different areas of European cybersecurity governance.

The methodological workflow is structured into three interconnected phases (Figure 1). First, a conceptual framing phase defines the Cyber-Social Security

perspective and identifies the protected environment as a socio-technical ecosystem rather than as a collection of isolated digital assets. The analysis considers the four foundational CSS domains – Cyber, Informational, Cognitive and Physical – and the interdependencies between technologies, people, organizations, information flows, decision-making processes and societal functions. Second, a propagation-oriented analysis extends the traditional Attack Surface through the notions of Influence Surface and Propagation Surface, allowing the analysis to move from the initial compromise to the trajectory of effects across domains.

Third, a governance and law mapping phase examines these cross-domain dynamics against selected components of the European regulatory ecosystem: the NIS2 Directive, the Critical Entities Resilience Directive (CER), the Digital Services Act (DSA), the Artificial Intelligence Act (AI Act), the Cyber Resilience Act (CRA), and the Cyber Solidarity Act (European Parliament and Council, 2025). These instruments address different objects, actors and responsibilities, but collectively cover several dimensions relevant to cyber-social risk, including cybersecurity risk management, critical-entity resilience, systemic platform risks, AI-enabled manipulation and transparency, product cybersecurity, vulnerability management, and coordinated preparedness and response.

Within this mapping, the Cyber Resilience Act – Regulation (EU) 2024/2847 (European Parliament and Council, 2024b) – plays a specific role at the product and technical-security layer. It establishes horizontal cybersecurity requirements for products with digital elements and introduces lifecycle-oriented obligations concerning secure design and development, cybersecurity risk assessment, vulnerability handling, security updates and responsibilities for manufacturers and other economic operators. From a CSS perspective, the CRA is relevant because products with digital elements may constitute direct or indirect entry points into interconnected socio-technical environments. Therefore, Secure-by-design and vulnerability-handling requirements can be interpreted as preventive barriers that reduce the probability that a product-level weakness becomes the trigger of a broader Cyber-Social cascade. CSS extends the analysis beyond product security by examining how a successful technical compromise may subsequently affect information integrity, decision autonomy, organizational operations and systemic resilience.

The resulting mapping is intentionally analytical rather than normative. CSS does not determine legal compliance, establish liability, or replace the sector-specific risk assessments required by European legislation, and the four CSS domains should not be interpreted as legal categories. The final output is a cross-domain governance perspective connecting cybersecurity, information integrity, decision autonomy, product security, systemic resilience and European law, with particular attention to propagation pathways and governance intervention points.

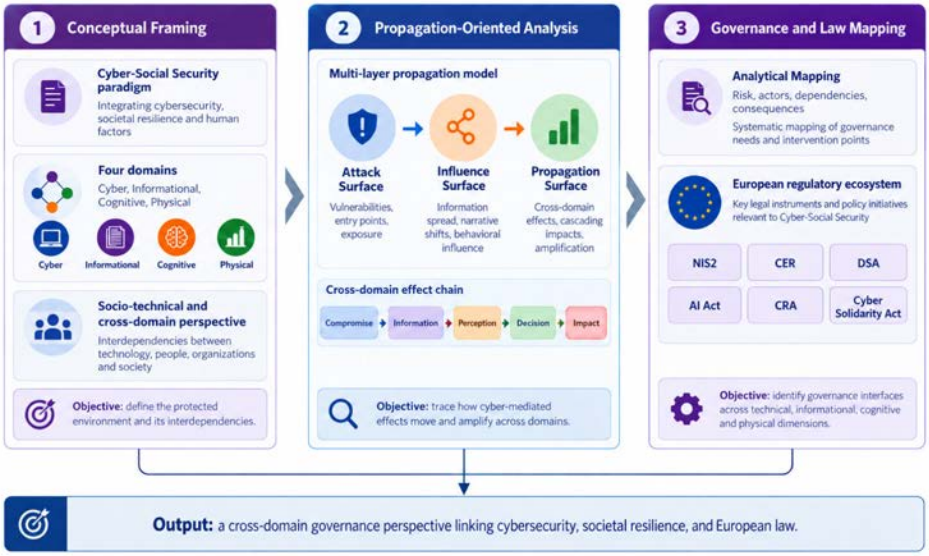

Fig. 1: CSS Methodological Approach: Interdisciplinary conceptual workflow for Cyber-Social Security governance analysis

3. From Cybersecurity to Cyber-Social Security

Traditional information security models have historically provided an effective foundation for protecting digital assets. The Confidentiality, Integrity and Availability triad remains fundamental for assessing whether data are protected from unauthorized disclosure and alteration and whether systems and services remain accessible. CVSS (Common Vulnerability Scoring System) and related vulnerability-management methodologies similarly provide essential mechanisms for comparing and prioritizing technical vulnerabilities (FIRST.Org, Inc., 2023; Bolpagni, 2022).

Therefore, the problem is not that these models are ineffective within their intended scope. The problem arises when a technical security metric is implicitly used as a proxy for the overall severity of a socio-technical event (Spring et al., 2021; Spring et al., 2021). A technically modest incident can produce severe consequences if it affects trusted information, influences decision-makers or disrupts a critical social process. Conversely, an incident receiving a very high technical severity score may remain operationally bounded if effective barriers prevent its effects from propagating beyond the affected technical environment (Barletta et al., 2024b).

The research underlying this contribution distinguishes precisely between latent socio-technical susceptibility and realized cross-domain event severity.

CSS addresses this gap by reorganizing the security environment around four interacting domains. The *Cyber domain* comprises infrastructures, software, networks, platforms, communication services, devices, algorithms and digital identities. It is frequently the entry point or amplification infrastructure of a cyber-social threat. The *Informational domain* concerns the production, transmission, authenticity, provenance, interpretation and contextualization of information. Attacks in this domain include fabricated or manipulated content, impersonation, deceptive narratives, information laundering and other mechanisms capable of degrading the quality of the information environment. The *Cognitive domain* concerns perception, beliefs, emotions, trust, reasoning and decision-making. Its inclusion is particularly important because the ultimate objective of an attack may not be to alter a file or disable a service but to induce a person, organization or community to behave differently. Finally, the *Physical domain* contains the tangible consequences for individuals, organizations, infrastructures, operational processes, urban environments and essential services.

These domains are analytically distinguishable but operationally interdependent (Figure 2). A synthetic video, for example, is generated and distributed using cyber infrastructure. Its credibility and contextual framing determine its effect within the informational domain. Exposure may subsequently influence trust, perception and judgment within the cognitive domain (Metzger and Flanagin, 2023). A resulting decision, such as emergency mobilization, market behavior, operational interruption or physical intervention, can then produce consequences in the physical domain. Research on misinformation and automated influence similarly demonstrates the relevance of cognitive and informational mechanisms to security outcomes (Pennycook and Rand, 2021; Rajtmajer and Susser, 2020).

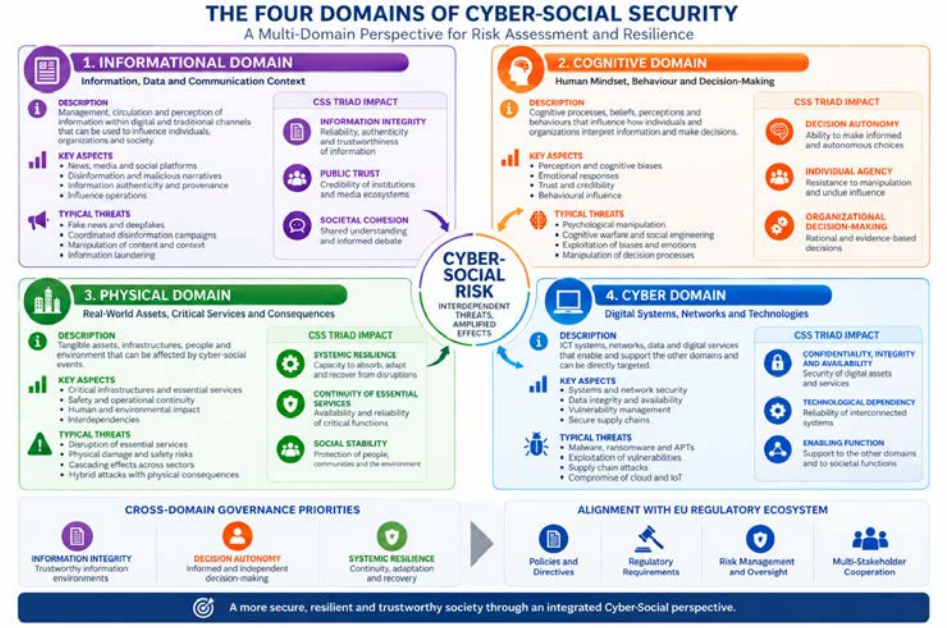

Fig. 2: The Four Domains of Cyber-Social Security and the CSS governance triad impact: Information Integrity, Decision Autonomy and Systemic Resilience.

The principal implication is that the unit of protection is no longer exclusively the digital asset; it is the socio-technical system in which that asset participates. This does not diminish traditional cybersecurity. Instead, it places technical security inside a larger causal and governance structure in which information quality, human judgment, organizational procedures and physical consequences can become security-relevant properties.

4. Attack Surface, Influence Surface, and Propagation Surface

A second conceptual change concerns how the security perimeter is represented. Traditional cybersecurity focuses heavily on the *Attack Surface*: the set of exposed assets, interfaces, services, identities, applications and vulnerabilities through which an adversary can obtain access or create a compromise. CSS broadens this view by considering not only where a threat originates, but also how the resulting effects interact with information, people, organizations and the physical environment. Therefore, the analytical focus moves from the compromised asset to the trajectory generated by the event.

Attack Surface – Where can the threat enter or generate the initial compromise? The Attack Surface identifies the point at which an adversary, malicious artefact or cyber-mediated action can introduce the initial perturbation into the socio-technical ecosystem. In addition to conventional technical vulnerabilities, this initial trigger may involve a compromised digital identity, a deepfake, a disinformation artefact, a social-engineering interaction or another mechanism capable of initiating a cyber-social trajectory. Its primary analytical focus is therefore the origin of the event.

Influence Surface – Who or what can be influenced? The Influence Surface represents the actors, groups, information channels, narratives, reputational assets, authority relationships and trust networks through which the initial event can modify perception, confidence, decision-making or behaviour. For example, a corporate email account belongs to the Attack Surface because it can be compromised; the authority associated with its owner, the credibility attributed to messages sent from that account and the organizational tendency to trust those communications belong to the Influence Surface. Consequently, this surface captures the capacity of an event to generate informational and cognitive effects and is closely related to the governance priorities of *Information Integrity* and *Decision Autonomy*.

Propagation Surface – Where can the effects travel, transform and amplify? The Propagation Surface is the set of technological connections, information flows, social relationships, organizational dependencies and environmental conditions that allow an effect generated in one domain to move into another. It shifts attention from the adversary' initial action to the evolution of the effect itself. Once initiated, propagation may continue even without further direct intervention by the attacker: weak source verification, delayed institutional communication, trusted relationships, algorithmic amplification or organizational procedures without adequate controls may facilitate cross-domain transitions and amplify the consequences of the original event.

Therefore, the three surfaces describe different but complementary moments of the same CSS trajectory (Figure 3):

- the *Attack Surface* identifies where the trajectory can begin;
- the *Influence Surface* identifies the people, information structures and trust relationships through which the event can shape perception and behaviour;
- the *Propagation Surface* identifies the dependencies and pathways through which those effects can move, change nature and produce further consequences.

A useful conceptual trajectory is *Compromise*, *Information*, *Perception*, *Decision*, *Action*, *Impact*. This sequence is illustrative rather than deterministic. In

Cyber-Social Security, effects may branch, converge, interact with other events or return to domains that have already been affected. A physical consequence can become new information, be reframed through digital platforms, alter perception and trigger a further organizational or cyber response.

The relationship among CSS domains can therefore assume a potentially circular form: *Cyber*, *Informational*, *Cognitive*, *Physical*, *Informational*, *Cognitive*, *Cyber*. A cyber-social attack should consequently be interpreted as a cascading and recursive phenomenon characterized by feedback loops. Its overall impact does not necessarily correspond to the simple sum of isolated domain impacts, because interactions among domains may progressively amplify, transform or attenuate the consequences of the initiating event. Therefore, a technically limited incident may produce significant societal effects, while a technically sophisticated attack may remain contained when cross-domain transitions are detected early and the propagation trajectory is interrupted.

To better illustrate the dynamics of cross-domain effect propagation, the wave metaphor can be introduced as a conceptual analogy within the Cyber-Social Security perspective (De Toni and Rullani, 2018). An initial cyberattack, deepfake or disinformation campaign can be understood as a perturbation (Terp and Breuer, 2022). Its eventual magnitude depends not only on the original event but also on the medium through which it travels. Social networks, recommender systems, cognitive biases, institutional distrust and organizational dependencies can act as amplification factors; technical controls, source verification, awareness, decision procedures and resilience mechanisms can attenuate or interrupt the trajectory (Figure 3).

This model has an important implication for governance: risk management should not be limited to preventing compromise. Where complete prevention is impossible, governance should seek interruption points along the propagation trajectory. A manipulated communication could therefore be mitigated technically through authentication controls, informationally through provenance verification, cognitively through awareness and independent verification procedures, organizationally through multi-party approval, and physically through continuity and emergency-response mechanisms. This changes the underlying security objective from asset protection alone to effect containment.

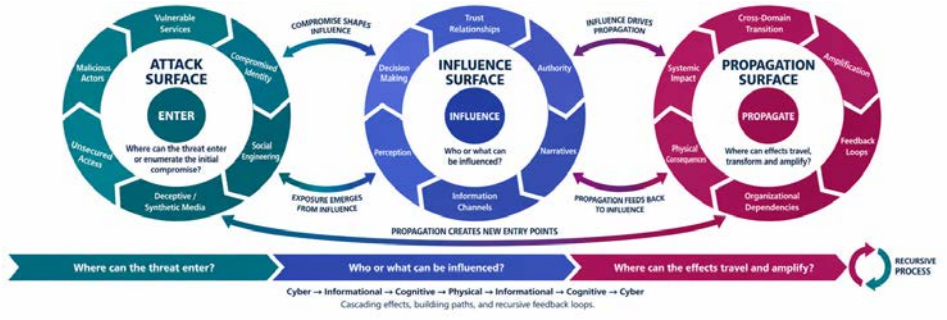

Fig. 3: Attack, Influence, and Propagation Surfaces in Cyber-Social Security. The model shows how a cyber-social event evolves from an initial entry point to influence processes and cross-domain propagation, highlighting amplification, feedback loops, and the recursive nature of effects across interconnected socio-technical domains.

5. Extending Cybersecurity Metrics: Exposure, Propagation and Realized Impact

The expansion from technical assets to socio-technical propagation creates a corresponding measurement problem. Confidentiality, Integrity and Availability continue to represent essential properties of information systems, but they do not directly capture information quality in context, manipulation of human judgment, or the capacity of socio-technical systems to absorb cascading disruption (Spring et al., 2021).

For this reason, the Cyber-Social Security perspective adopted in this article emphasizes three broader governance priorities: Information Integrity, Decision Autonomy, and Systemic Resilience. These categories translate the underlying cyber-social framework into a vocabulary better suited to policy analysis, organizational governance, and legal interpretation.

Information Integrity refers to the reliability, provenance, traceability, contextual consistency, and trustworthiness of the information environment. It goes beyond the narrow technical meaning of data integrity. Information may remain technically unaltered while being decontextualized, reframed, selectively amplified, falsely attributed, or embedded within manipulative narratives. From a governance perspective, protecting information integrity means safeguarding the conditions under which individuals, organizations, and institutions can rely on information for meaningful coordination and accountability.

Decision Autonomy concerns the ability of individuals and organizations to make informed choices without being subject to covert, deceptive, or manipulative forms of influence. This priority does not imply the absence of persuasion

or social influence, which are intrinsic to social life. Rather, it concerns the preservation of meaningful autonomy in environments shaped by algorithmic amplification, synthetic media, behavioral targeting, and information asymmetries. Within the CSS perspective, decision autonomy becomes a central governance concern because many cyber-social attacks seek not merely to compromise systems but to steer decisions, reactions, and behavior.

Systemic Resilience refers to the capacity of socio-technical systems, institutions, and critical services to absorb, adapt to, and recover from disruptions without allowing localized digital shocks to become broader societal crises. It includes organizational continuity, redundancy, institutional preparedness, crisis coordination, and the ability to restore operational functionality and trust after a disruptive event. In this sense, systemic resilience exceeds the traditional meaning of technical recovery and incorporates the stability of interdependent infrastructures and social functions.

These three governance priorities distinguish different layers of harm and can be applied both before and after an incident. Pre-incident analysis examines structural susceptibility: weaknesses in infrastructure, information-validation processes, trust relationships, organizational controls, human-system interaction and cyber-physical dependencies. Post-incident analysis measures realized effects once a threat has been operationalized. In this lifecycle, the Cyber-Social Vulnerability Scoring System (C-SVSS) can be interpreted as a pre-incident representation of latent structural exposure, while Cyber-Social Event Severity (CSES) represents the post-incident severity of a realized event and its cross-domain consequences.

This perspective also clarifies the difference between pre-incident exposure and post-incident consequences. In the pre-incident phase, C-SVSS can be interpreted as a structured representation of latent cyber-social exposure. It brings together technical vulnerabilities, weaknesses in the information environment, human and organizational susceptibility, interdependencies and contextual criticality. The output may be expressed as an ordinal severity class or a normalized score, but the emphasis in this paper is on the dimensions being considered rather than on a specific calculation rule.

In the post-incident phase, CSES shifts attention from potential exposure to realized event severity. It considers the sophistication of the initiating event, the degree of cross-domain propagation and the observed consequences for Information Integrity, Decision Autonomy and Systemic Resilience. Figure 4 summarizes this pre-incident/post-incident lifecycle and separates structural exposure, the triggering threat, propagation and realized impact.

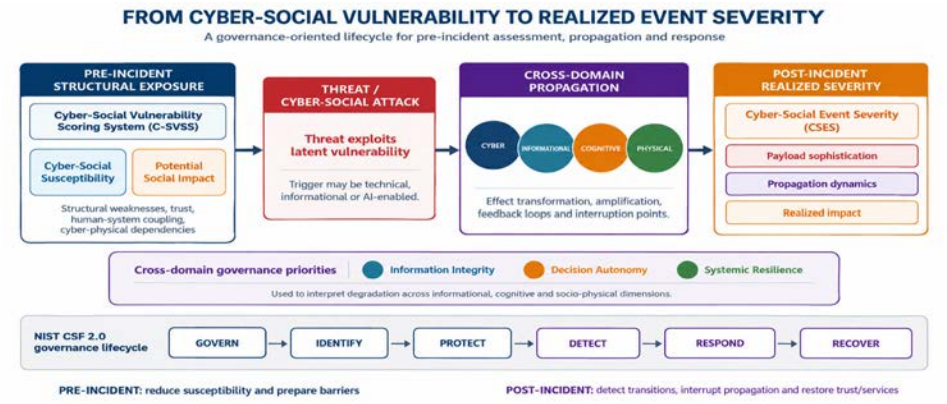

Fig. 4: *From Cyber-Social Vulnerability to Realized Event Severity: a governance-oriented lifecycle. C-SVSS represents pre-incident structural exposure, the attack triggers cross-domain propagation, and CSES characterizes realized severity. The lifecycle is aligned conceptually with NIST CSF 2.0 functions*

The distinction aligns well with the logic of NIST CSF 2.0 (National Institute of Standards and Technology, 2024). Pre-incident analysis corresponds primarily to the Govern, Identify, and Protect functions, while post-incident analysis becomes more relevant to Detect, Respond, and Recover. From this perspective, Cyber-Social governance is not limited to preventing compromise; it also seeks to detect transitions, interrupt propagation, contain consequences, and restore trust and functionality once the event has unfolded. Therefore, the resulting metric logic can be summarized as Exposure, Propagation, Realized Impact.

Figure 5 visually captures a fundamental proposition: Cyber-Social risk depends not merely on the intensity of the initiating event, but on the interaction between system susceptibility, possible impact and cross-domain escalation. A highly sophisticated attack can generate limited systemic risk when propagation is successfully contained. Conversely, relatively simple manipulation can produce significant consequences in highly exposed socio-digital environments. The figure also makes explicit that European regulatory instruments operate as governance interfaces around the multi-domain risk space rather than as direct equivalents of CSS domains.

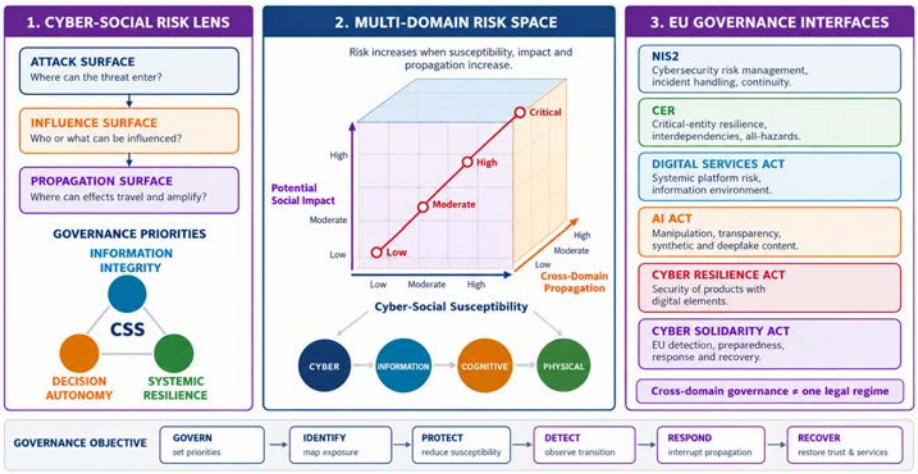

Fig. 5: Cyber-Social Security × EU Regulatory Ecosystem: a cross-domain governance perspective linking cyber-social susceptibility, potential social impact and effect propagation with the main European regulatory interfaces.

6. Agentic AI, LLMs and Human Oversight in Cyber-Social Governance

Artificial Intelligence (AI) performs a dual role within the CSS landscape. It can constitute an attack-enabling technology through synthetic media, automated impersonation, personalization, or scalable influence but it can also support detection, analysis, and incident response. This duality becomes more significant as the analytical role of AI evolves from isolated model inference toward coordinated, multi-step decision-support workflows.

Large Language Models (LLMs) provide a semantic reasoning layer that can interpret unstructured reports, forensic artefacts, OSINT observations, regulatory text and heterogeneous security signals (Alqahtani and Kumar, 2026). Agentic AI extends this capability by adding goal decomposition, tool selection, state management and sequential orchestration. In a Cyber-Social Security context, an agentic workflow can therefore decompose an investigation into bounded tasks, invoke specialized tools, compare evidence across domains, update risk variables, and propose mitigation actions as the threat trajectory evolves.

This research work does not treat Agentic AI as an autonomous governance authority. Instead, it is positioned as an orchestration and decision-support layer operating across CSS lifecycle. Before an incident, agents may support Govern, Identify and Protect by gathering exposure indicators and organizing C-SVSS related evidence. During an event, they may support Detect by correlating changes across the Cyber, Informational, Cognitive and Physical domains. Instead, during response

and recovery, they may compare candidate interventions, highlight expected benefits and risks, and present recommendations to authorized human decision-makers.

A possible implementation substrate for this model is a Digital Twin (DT)-based CSS environment. In such a conceptual architecture, the DT represents the evolving socio-technical state, LLMs provide semantic interpretation of heterogeneous evidence, and bounded agents coordinate analytical tasks such as log triage, provenance checking, taxonomy mapping, anomaly correlation, simulation and preliminary report generation (Barletta et al., 2026). Therefore, the DT acts as a decision-support environment rather than as an autonomous decision-maker, while the agentic layer provides the orchestration required to move from static analysis toward adaptive cross-domain assessment.

The resulting architecture can be understood as a progression from AI-assisted analysis to agentic Cyber-Social decision support: LLMs interpret heterogeneous evidence; specialized agents organize and cross-check analytical tasks; CSS indicators structure exposure, propagation and impact information; and human decision-makers validate the interpretation and authorize material interventions. This separation is particularly important in governance contexts because semantic plausibility, operational urgency and legal permissibility are different questions and should not be collapsed into a single automated output.

Figure 6 provides a functional view of the proposed Cyber-Social decision-support architecture, emphasizing the interaction among analytical components rather than the internal operation of individual AI techniques. Heterogeneous evidence originating from cyber telemetry, OSINT, media forensics, provenance analysis, operational data, and governance sources enters a common analytical pipeline. The semantic layer transforms these inputs into contextualized evidence, which is subsequently distributed across specialized analytical functions coordinated by the agentic orchestration layer. The Digital Twin maintains a dynamic representation of the socio-technical environment and provides a shared state on which these functions can operate and perform scenario-based analysis. The resulting evidence is then organized through the three CSS analytical perspectives: the Attack Surface identifies the initial exposure or entry point, the Influence Surface captures the actors, trust relationships, and decision processes that may be affected, and the Propagation Surface represents the pathways through which effects may spread, transform, or amplify. These observations are correlated across the Cyber, Informational, Cognitive, and Physical domains and translated into a governance-oriented assessment combining pre-incident exposure, cross-domain propagation, and realized event severity. Human oversight constitutes the final validation layer, ensuring that mitigation, response, recovery, and knowledge-update actions remain subject to authorized interpretation and approval. The feedback path closes the architecture by allowing newly observed evidence and post-response information to update the evolving representation of the Cyber-Social environment.

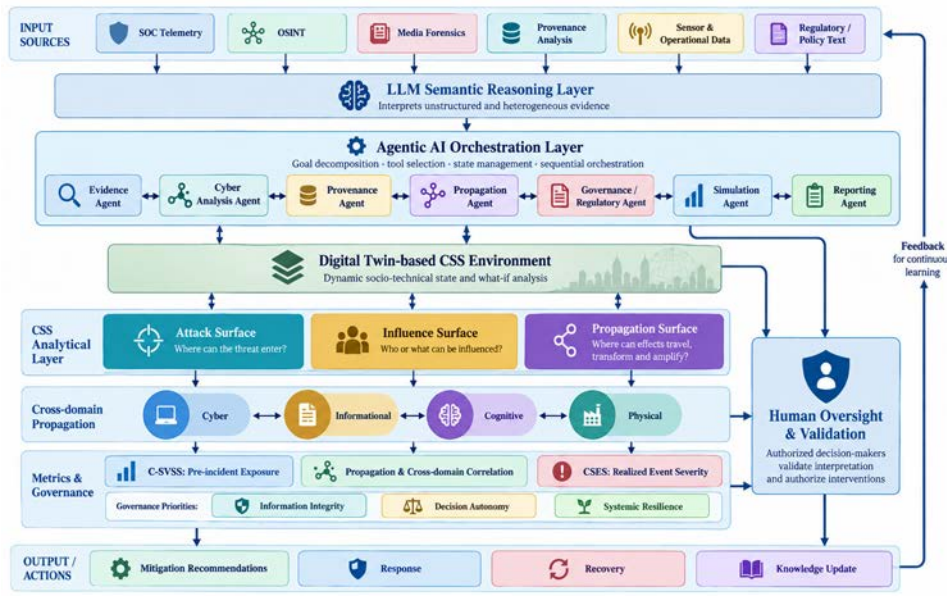

Fig. 6: Agentic AI, LLMs and Human Oversight in Cyber-Social Governance.

7. Cyber-Social Security and The European Regulatory Ecosystem

The European digital regulatory framework offers a particularly useful environment in which to examine the governance implications of CSS. Recent EU legislation does not establish a single unified concept of cyber-social risk. Instead, different legal instruments address different portions of the socio-digital ecosystem. From a CSS perspective, their interaction is therefore more relevant than any attempt to identify a one-to-one correspondence between a regulation and a domain.

The NIS2 Directive remains strongly linked to cybersecurity but adopts a broader risk-management logic than a purely technical vulnerability framework. Article 21 requires appropriate and proportionate technical, operational and organizational measures, including risk analysis, incident handling, business continuity, crisis management, supply-chain security, vulnerability handling, cybersecurity hygiene, training and human-resource security (European Parliament and Council, 2022a). This is compatible with the CSS proposition that organizational dependencies and human processes form part of the security environment.

The Critical Entities Resilience Directive further reinforces this systemic shift. It requires attention to natural and man-made risks, cross-sectoral and cross-border dependencies, hybrid threats and the ability of critical entities to prevent, protect against, respond to, resist, mitigate, absorb, accommodate and recover from incidents. The Directive also emphasizes coordination be-

tween authorities responsible for critical-entity resilience and those operating under NIS2. The resulting governance model is particularly relevant to the CSS Physical domain because the physical consequence of a digital event frequently emerges through dependencies between infrastructures, organizations and essential services rather than through the compromised asset alone.

The Digital Services Act moves the legal analysis further into the informational and social environment. Article 34 requires providers of very large online platforms and very large online search engines to identify, analyse and assess systemic risks arising from the design or functioning of their services, including algorithmic systems. The specified risks extend beyond illegal content to fundamental rights, civic discourse and electoral processes, public security, gender-based violence, public health, minors, and serious consequences for physical and mental well-being (European Parliament and Council, 2022c). From a CSS perspective, the importance of this provision lies in its recognition that digital architectures and algorithmic systems can create or amplify consequences across informational, cognitive and societal dimensions.

The AI Act provides an additional bridge between technological governance and human agency. Its prohibition of specific manipulative or deceptive practices explicitly addresses circumstances in which AI materially distorts behavior and creates or is reasonably likely to create significant harm. Its transparency regime is also directly relevant to information authenticity: Article 50 establishes obligations for specified AI-generated or manipulated content and for deepfakes in the circumstances defined by the Regulation (Calvano et al., 2025; European Parliament and Council, 2024a; European Commission, 2026).

The Cyber Resilience Act complements this environment from the product-security perspective by establishing cybersecurity requirements for products with digital elements (European Parliament and Council, 2024b). Within CSS, such obligations are primarily located at the cyber/technical layer but can also be understood as preventive mechanisms intended to reduce the probability that technical weaknesses become triggers for wider propagation.

Finally, the Cyber Solidarity Act reinforces Union-level capabilities for coordinated threat detection, preparedness and response. It establishes mechanisms aimed at strengthening European capacities to detect cyber threats and incidents and to prepare for and respond to significant and large-scale incidents (European Parliament and Council, 2025). The common situational-awareness component is especially relevant to CSS because effective management of cross-domain events requires the ability to correlate heterogeneous signals rather than observe technical incidents in isolation.

Table 1 summarizes the mapping between CSS dimensions and selected European governance instruments and, as a result, it reveals governance interfaces that may otherwise remain separated.

Table 1. Functional mapping between CSS dimensions and selected European governance instruments

CSS dimension	Predominant governance concern	Examples of EU legal relevance
Cyber	Security of networks, systems, products and infrastructures	NIS2; Cyber Resilience Act; Cyber Solidarity Act
Informational	Authenticity, provenance, content integrity and systemic information risks	Digital Services Act; AI Act
Cognitive	Human autonomy, manipulation, trust and informed decision-making	AI Act; Digital Services Act
Physical	Continuity, essential services, infrastructure and societal resilience	CER Directive; NIS2
Cross-domain propagation	Interdependency, systemic risk, crisis coordination and impact containment	NIS2–CER coordination; DSA systemic-risk governance; Cyber Solidarity Act

Consider an AI-generated impersonation used against an operator of essential services. The generation and dissemination infrastructure represents a Cyber dimension. The apparent authenticity and framing of the communication belong to the Informational domain. The induced belief and resulting operator decision belong to the Cognitive domain. If the action disrupts an essential service, the effects reach the Physical domain. Depending on the facts, different components of this trajectory may become relevant to AI transparency, platform governance, cybersecurity risk management, incident handling, resilience or other legal regimes.

The CSS perspective therefore adds a question that traditional compliance mapping often does not foreground: *How can an incident move from one regulated environment into another?* This is where the concept of Propagation Surface becomes relevant to law and governance. Governance should account not only for the obligations applicable at the initial point of compromise but also for the regulatory and institutional interfaces encountered as effects move across systems, actors and domains.

8. Conclusion

The increasing convergence of cyberspace and society challenges the assumption that cybersecurity incidents can be assessed solely by examining compromised systems, software vulnerabilities and technical properties. Contemporary cyber-mediated threats can alter information environments, influence perception and decision-making, exploit social and organizational relationships and ultimately generate tangible consequences for people, institutions and critical infrastructures.

Cyber-Social Security responds to this transformation by interpreting security across four interconnected domains (Cyber, Informational, Cognitive and Physical) and by shifting attention from individual domains to the relationships among them. The resulting perspective complements the traditional Attack Surface with an Influence Surface and a Propagation Surface, making the trajectory of an effect itself an object of security analysis.

This shift has direct implications for governance. The objective is no longer exclusively to prevent an adversary from entering a system, but also to identify where an effect can be amplified, transformed or interrupted. In this sense, Cyber-Social resilience can be interpreted as the capacity not only to withstand a cyberattack but also to prevent a localized compromise from becoming an informational, cognitive, organizational or physical crisis.

The European regulatory ecosystem provides a particularly significant context for this development. NIS2 governs cybersecurity risk management and organizational resilience; CER focuses on the resilience and interdependencies of critical entities; the DSA targets systemic risks associated with digital platforms; the AI Act regulates manipulation and transparency in AI-mediated environments; the Cyber Resilience Act establishes requirements for the security of digital products; and the Cyber Solidarity Act strengthens coordinated detection, preparedness, and response capabilities. None of these instruments is equivalent to the CSS paradigm, and CSS does not replace their respective compliance frameworks. Taken together, however, they demonstrate the growing policy relevance of risks that cannot be confined to a single technical perimeter.

Therefore, the principal contribution of Cyber-Social Security is not the creation of another isolated security layer. Its contribution lies in providing a common conceptual perspective through which technical compromise, information authenticity, cognitive autonomy, human oversight and societal resilience can be examined as parts of the same evolving risk trajectory.

As a result, the fundamental question of cybersecurity is changing. It is no longer only "Has the system been compromised, and with what technical impact?", it becomes: "Where does the threat originate, through which domains can its effects propagate, how can those effects influence people, organizations and infrastructures, and where can governance intervene to prevent a local incident from becoming a systemic crisis?".

References

Agrafiotis, I., Nurse, J. R. C., Goldsmith, M., Creese, S., Upton, D. 2018. A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity*, 4(1), tyy006. https://doi.org/10.1093/cybsec/tyy006.

Alqahtani, H., Kumar, G. 2026. Large Language Models for Cybersecurity Intelligence: A systematic review of emerging threats, defensive capabilities, and security evaluation frameworks. *Computers, Materials & Continua*, 87(3). https://doi.org/10.32604/cmc.2026.077367.

Antoniol, V., Barletta, V. S., Battista, F., Buono, P., Caivano, D., Campesi, G., Cascione, G., Curci, A., de Gemmis, M., Gattulli, V., Scardigno, R., Sciacovelli, A. L., Sorianello, P., Tamburrano, V. 2025. Security and Rights in CyberSpace: Cyber Social Security. In *Proceedings of the Workshop on Collaboration and Learning through Symbiotic Artificial Intelligence (COL-SAI 2025), co-located with the 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025)*, CEUR Workshop Proceedings, Vol. 4086, pp. 25–33.

Austin, G., Withers, G. 2020. Creating social cyber value as the broader goal. In Austin, G. (Ed.), *Cyber Security Education* (Vol. I, 1st ed.), pp. 99–118. London: Routledge. https://doi.org/10.4324/9780367822576-6.

Barletta, V. S., Caivano, D., Catalano, C., de Gemmis, M., Impedovo, D. 2024a. Cyber Social Security Education. In De Paolis, L. T., Arpaia, P., Sacco, M. (Eds.), *Extended Reality. XR Salento 2024*, Lecture Notes in Computer Science, Vol. 15030, pp. 240–248. Cham: Springer. https://doi.org/10.1007/978-3-031-71713-0_16.

Barletta, V. S., Calvano, M., Sciacovelli, A.L. 2024b. Cyber Social Security in Multi-Domain Operations. In *Proceedings of the 2024 IEEE International Workshop on Technologies for Defense and Security (TechDefense)*, Naples, Italy, pp. 41–46. IEEE. https://doi.org/10.1109/TechDefense63521.2024.10863352

Barletta, V. S., Bavaro, V., Calvano, M., Curci, A., Piccinno, A., Posa, D. P. 2026. Enabling Cybersecurity Education Through Digital Twins and Generative AI. In Maciel, C., Spano, D. L., Ardito, C., Freire, A., Prates, R. (Eds.), *Blending Experiences in Interaction Design. INTERACT 2025*, Lecture Notes in Computer Science, Vol. 16478. Cham: Springer. https://doi.org/10.1007/978-3-032-22934-2_17.

Bolpagni, M. 2022. Cyber risk index: A socio-technical composite index for assessing risk of cyber attacks with negative outcome. *Quality & Quantity*, 56, 1643–1659. https://doi.org/10.1007/s11135-021-01199-3.

Calvano, M., Curci, A., Lanzilotti, R., Piccinno, A. 2025. Guiding Designers and Developers in Creating Symbiotic AI Systems. In *Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly '25)*, Article 62, pp. 1–5. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3750069.3750095.

Costa, A., Crupi, A., Cesaroni, F., Abbate, T. 2025. Exploring the role of artificial intelligence in addressing sustainable development. A semantic analysis of AI patents. *Technovation*, 148, 103335. https://doi.org/10.1016/j.technovation.2025.103335.

De Toni, A. F., Rullani, E. (Eds.). 2018. *Uomini 4.0: ritorno al futuro. Creare valore esplorando la complessità*. Milano: FrancoAngeli.

European Commission. 2026. *Guidelines on the transparency obligations for providers*

and deployers of certain AI systems under Article 50 of Regulation (EU) 2024/1689. Brussels: European Commission.

European Parliament and Council. 2022a. Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS2 Directive). *Official Journal of the European Union*, L 333.

European Parliament and Council. 2022b. Directive (EU) 2022/2557 of 14 December 2022 on the resilience of critical entities and repealing Council Directive 2008/114/EC (CER Directive). *Official Journal of the European Union*, L 333.

European Parliament and Council. 2022c. Regulation (EU) 2022/2065 of 19 October 2022 on a Single Market for Digital Services (Digital Services Act). *Official Journal of the European Union*, L 277.

European Parliament and Council. 2024a. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

European Parliament and Council. 2024b. Regulation (EU) 2024/2847 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act). *Official Journal of the European Union*.

European Parliament and Council. 2025. Regulation (EU) 2025/38 laying down measures to strengthen solidarity and capacities in the Union to detect, prepare for and respond to cyber threats and incidents (Cyber Solidarity Act). *Official Journal of the European Union*.

FIRST.Org, Inc. 2023. *Common Vulnerability Scoring System Version 4.0: Specification Document* (Version 1.2). FIRST.Org, Inc.

Metzger, M. J., Flanagin, A. J. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59(B), 210–220. https://doi.org/10.1016/j.pragma.2013.07.012.

National Institute of Standards and Technology. 2024. *The NIST Cybersecurity Framework (CSF) 2.0*. NIST CSWP 29. Gaithersburg, MD: National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.29.

Pennycook, G., Rand, D. G. 2021. The psychology of fake news. *Trends in Cognitive Sciences*, 25, 388–402. https://doi.org/10.1016/j.tics.2021.02.007.

Rajtmajer, S., Susser, D. 2020. Automated influence and the challenge of cognitive security. In *Proceedings of the 7th Symposium on Hot Topics in the Science of Security (HotSoS '20)*, pp. 62–70. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3384217.3385615.

Spring, J. M., Hatleback, E., Householder, A. D., Manion, A., Shick, D. 2021. Time to change the CVSS? *IEEE Security & Privacy*, 19(2), 74–78. https://doi.org/10.1109/MSEC.2020.3044475.

Terp, S. J., Breuer, P. 2022. DISARM: A framework for analysis of disinformation campaigns. In *Proceedings of the 2022 IEEE International Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pp. 1–8. IEEE. https://doi.org/10.1109/CogSIMA54611.2022.9830669.

Zegarow, P., Bartuzi, E. 2024. Deepfake influence tactics through the lens of Cialdini's principles: Case studies and the DEEP FRAME tool proposal. *Applied Cybersecurity & Internet Governance*, 3(2), 286–302. https://doi.org/10.60097/ACIG/201147

AT THE HEART OF A CYBER CRISIS: THE KEY TO SUCCESS

Dominique Kogue
LHC – Luxembourg House of Cybersecurity
National Cybersecurity Competence Center, Luxembourg

Abstract
Cybersecurity is no longer just about preventing attacks. Despite the right technologies, security measures, and procedures, incidents can still occur and escalate into crises that disrupt business operations. In such situations, having a plan is essential, but knowing how to put it into practice effectively is what truly makes the difference. This chapter explores how to prepare for and manage a cyber crisis. Drawing on insights from ROOM#42 cyber crisis exercises, it examines the challenges of responding to unexpected situations, coordinating actions, and making decisions with incomplete information. These exercises show that effective crisis management requires more than technical expertise. Artificial intelligence can help security teams analyse information and respond faster, but it can also make cyberattacks more convincing and accelerate crises. As technology evolves, human judgment remains essential. The goal is not to prevent every crisis, but to be better prepared to face the unexpected, make sound decisions under pressure, and learn from experience to build resilience.

1. Understanding the Cyber Crisis

A cybersecurity incident does not necessarily become a crisis. Security events occur daily, and many can be managed through established operational processes as highlighted in the ENISA Threat Landscape 2026. An employee may receive a suspicious email, a security system may detect unusual activity, or a workstation may become infected with malware. In most cases, the security or IT team investigates the event, contains the threat, restores normal operations, and moves on. A crisis begins when these normal processes are no longer sufficient to manage the situation and its consequences. This escalation may result from a significant disruption to business operations or those of external partners, potential regulatory fines or sanctions, financial losses, or reputational damage. At this stage, cybersecurity is no longer merely a technical or IT concern; it becomes a business risk, a management challenge, and ultimately a matter of leadership.

The situation may develop gradually, but can also escalate into a crisis immediately, depending on the nature and impact of the event. An unusual event may initially appear to be an isolated technical problem. Then additional systems become unavailable, employees report difficulties, end users or customers

begin to complain, or information appears to have been stolen. Management starts asking questions that the security team cannot immediately answer. Legal advisers become involved because there may be regulatory or contractual consequences. Communications teams begin to consider whether employees, customers or the public need to be informed. What started as a technical incident has progressively become a crisis.

The most difficult characteristic of a cyber crisis is not necessarily the complexity of the technology. It is the combination of uncertainty and time pressure. During normal operations, decision-makers can often take the time to collect information, compare alternatives and seek approval. During a crisis, waiting for complete information may itself create additional damage. While the teams are still trying to understand what has happened, an attacker may be moving through the network, stealing information, disabling systems or preparing another stage of the attack. Critical decisions must therefore be made and action taken while the full extent and implications of the situation remain unclear.

This changes how decisions are made. The goal is not to wait until everything is known before making the perfect decision. The goal is to make a good enough decision quickly, protect the organisation, and preserve future options. A different mindset is required, where Leaders must be comfortable with uncertainty while remaining disciplined. They need to separate what is known, what is suspected, and what is still unknown. They must also decide which actions cannot wait and which can be delayed until more information is available.

This is where the importance of people becomes clear. Cybersecurity is often described as a technological discipline, but crisis management demonstrates its human dimension very clearly. Security technologies can detect anomalies, block malicious activity, and provide information. They cannot, by themselves, decide whether an operation or a service should be shut down, activate business continuity procedures, contact a regulator, communicate with customers or request external assistance. These decisions require judgment, context and responsibility.

The crisis cell must include representatives from all key operational functions. A security specialist may understand the technical characteristics of an attack but may not know which business processes are most critical. A business manager may understand the operational consequences but not know whether a particular system has been compromised. Legal advisers may understand the regulatory exposure, while communications specialists understand how employees, customers, and the public may perceive the information. No single perspective is sufficient. An effective crisis response requires bringing these perspectives together so decisions reflect the situation's full impact. Technical expertise alone is not enough to manage a crisis. Effective crisis management also requires clear communication, prioritisation, coordination, escalation, and the ability to make fast, informed decisions under pressure.

Communication itself can become a second crisis. During a cyberattack, information is fragmented. The security team knows some things, business teams know others, employees see symptoms, customers experience disruption, and the attacker may deliberately spread misinformation. Control over the information environment can therefore be lost at the same time as efforts are made to regain control of the technology. This is particularly important in an environment where social media and AI-generated content can spread convincing information very quickly.

It is therefore essential to communicate honestly and not pretend to know more than we actually do. There is an important difference between saying, "We know this happened," "We believe this is happening," and "We are investigating whether this happened." Keeping these distinctions clear is essential for maintaining credibility. During a crisis, transparency does not mean sharing everything immediately. It means clearly communicating what is known, explaining what is still being investigated, and avoiding speculation that could create further uncertainty.

Artificial intelligence adds another dimension to this environment. It can help defensive teams work faster by processing large amounts of information, connecting events, and identifying patterns. At the same time, attackers can use AI to improve reconnaissance, create convincing messages, images, and videos, and automate parts of social engineering. As a result, both defenders and attackers can operate faster, making the threat environment more dynamic and difficult to manage, which creates an important paradox for the future of cyber crisis management. The more capable machines become, the more important human judgment becomes for decisions that should not be delegated blindly. AI may be able to tell what is happening, recommend a response, or even execute predefined actions. But ultimately, the decision remains with people. Management and security experts must decide how much risk the organization is willing to accept and who will be accountable for the consequences.

For this reason, three themes provide a useful foundation for understanding cyber crisis management: Time, People and Skills. Time determines how quickly action is needed. People determine how information is understood, decisions are made, and efforts are coordinated. Skills determine whether people can perform effectively under pressure. These three dimensions are closely connected. A lack of skills can waste valuable time. Poor coordination can lead to delays and confusion. And when time is limited, the pressure on people and their ability to use their skills effectively increases.

The central lesson is simple: a cyber crisis is not just a more serious cybersecurity incident. It is a different kind of challenge. It requires people to make important decisions with incomplete information, often under intense time pressure, while bringing together different roles, responsibilities, and points

of view. Technology remains essential, but it cannot manage a crisis on its own. What matters just as much is having the right people, skills, structures, and decision-making processes in place so that technology can be used effectively when normal conditions no longer apply.

2. Preparing for the Crisis: Readiness and Resilience

Crisis management becomes much less stressful when people have prepared for it in advance. This may sound obvious, but preparedness is often misunderstood. Many organisations have incident response plans, business continuity procedures and emergency contact lists. They may even have policies approved by management and stored in a document repository. But having a plan on paper does not mean being ready to use it. Real preparedness means that people know what to do, who can make decisions, how information should be shared, and how operations can continue when normal ways of working are no longer possible.

A crisis is when you find out whether your plans actually work or are simply documents sitting on a shelf. If corporate email is unavailable, can the crisis team still communicate? If the person responsible for incident response cannot be reached, does someone else have the authority to take over? If a critical system needs to be isolated, does the team understand the impact this could have on the business? And if the attacker has compromised the identity infrastructure, can the crisis team still access the systems and information it needs? These questions should not be answered for the first time during a crisis. They should be addressed, tested, and understood beforehand. Good crisis management depends not only on having plans, but on knowing that those plans will work when they are needed most.

Preparedness starts with understanding what really needs to be protected. This means looking beyond individual applications or servers and identifying the services and activities that are essential to keep things running. Which functions must continue? Which processes depend on them? What information is critical? Which people are needed to keep them operating? And which external suppliers are involved? Answering these questions helps build a clear picture of how everything is connected. This understanding becomes extremely valuable during a crisis, when decisions often need to be made quickly, and the consequences of disrupting one part of the business can affect many others.

This perspective is important because modern businesses are deeply interconnected. A company may rely on cloud services, telecommunications providers, payment platforms, software vendors, managed-service providers, and external data centers. A disruption affecting one of these dependencies can quickly become a crisis for the company itself. Even with strong internal cyber-

security controls, a company may be unable to operate if a critical external service becomes unavailable. Resilience therefore requires looking beyond internal systems and understanding the wider ecosystem on which a company depends. This is also why recent regulations and directives such as NIS2, DORA and, to some extent, the Cyber Resilience Act (CRA) place greater emphasis on supply-chain and third-party risks. Managing these dependencies has become an important part of building and maintaining cyber resilience.

The next step is to establish a clear crisis-management structure. When an incident occurs, valuable time should not be lost trying to figure out who is responsible for coordinating the response. Everyone should already know who leads the crisis, who provides technical expertise, who represents the business, who manages communications, and who provides legal or regulatory advice. Backup people should also be identified. Effective crisis management cannot depend on one person being available at exactly the right moment.

Clear authority is just as important. During a crisis, delays often occur because people are unsure who has the authority to make decisions. Can the security team disconnect a server? Can IT disable an account belonging to a senior executive? Can business continuity measures be activated immediately? Can the communications team issue a public statement? Can external specialists be called in without going through several levels of approval? These decisions are much easier to make before a crisis occurs. Everyone should know what they are authorized to do, when they can act on their own, and when they need to involve someone else. Clear decision-making authority helps avoid unnecessary delays when every minute matters.

Communication deserves particular attention. Companies often rely on the same digital systems for both everyday work and crisis management. If corporate email, collaboration tools, or identity systems are compromised, the very tools needed to coordinate the response may no longer be available. For this reason, alternative communication methods should be identified and tested in advance. Emergency phone numbers, independent messaging systems, physical meeting locations, and offline contact lists can become essential when normal systems are unavailable. The ability to communicate should never depend entirely on the systems that may be affected by the crisis itself. But having an alternative is not enough. A communication channel that has never been tested is only an assumption. The same applies to backups, emergency procedures, and recovery processes. During a real incident, companies often discover that a backup exists but cannot be restored quickly enough, that an emergency contact is no longer valid, or that a supposedly independent communication channel still depends on the same compromised identity infrastructure. Preparedness therefore comes from testing these capabilities under realistic conditions and fixing the weaknesses before a crisis exposes them.

Evidence preservation is another area where preparation is important. During an attack, the natural reaction is to fix the problem as quickly as possible. But some actions taken to restore normal operations can also destroy valuable evidence about what happened. Logs may be overwritten, systems may be rebuilt, and suspicious files may disappear. This does not mean delaying action when an attack needs to be contained. It means considering containment and evidence preservation at the same time. Before a crisis occurs, it should be clear who is responsible for forensic matters, what evidence needs to be preserved, and when external specialists should be brought in. This allows the response team to act quickly without losing information that may later be essential for understanding the attack, meeting legal or regulatory requirements, or supporting further investigation.

External support should also be part of the preparedness plan. Many companies do not have all the cybersecurity capabilities they need in-house, and some may not have cybersecurity experts at all. It is therefore important to understand where the limits are and know when to bring in additional expertise. Depending on the situation, this could mean contacting a CERT or CSIRT, specialist incident responders, forensic experts, legal advisers, technology providers, or law enforcement. Asking for help should not be seen as a failure. In many cases, it shows that those responsible understand the seriousness of the situation and know when and how to use the wider cybersecurity ecosystem.

These limitations often only become clear when people are put to the test, which is why exercises are such an important part of preparedness. They show whether plans, roles, and procedures actually work when people have to make decisions under pressure. More importantly, they reveal weaknesses that may not be visible on paper: unclear responsibilities, poor communication, gaps in technical capabilities, or uncertainty about when to ask for external help. The real value of an exercise is not to prove that everything is ready. It is to identify what needs to improve. Each exercise provides an opportunity to learn, adjust procedures, and strengthen the ability to respond to unexpected situations. In this sense, preparedness is not a one-time activity. It is a continuous process of testing, learning, and improving. This is what gradually turns preparedness into resilience: the ability not only to respond to a crisis, but also to adapt when the situation does not go as planned.

Technology is an important part of resilience, but it is not enough on its own. An effective crisis response also depends on people, skills, processes, and clear decision-making working together. This brings us back to the three elements at the heart of preparedness: Time, People, and Skills. People need to know what to do, have the skills to do it, and act quickly. The goal of preparation is not to predict exactly what the next attack will look like, but to build the ability to respond and adapt when faced with something unexpected.

3. Preparedness: ROOM#42 an Immersive Experience Made in Luxembourg

There is a fundamental difference between reading about a crisis and experiencing it. A document can explain what should happen. A procedure can describe who is responsible. A policy can establish principles. But none of these creates the experience of making decisions while the situation is changing and the information available is incomplete. This is why exercises are such an important part of cyber resilience. To provide this kind of realistic and immersive experience, ROOM#42, a cyberattack and crisis simulator, was created in Luxembourg in 2018. Its objective was to put humans at the center of incident and crisis management and allow them to experience the pressure, uncertainty, and decision-making challenges of a real cyber crisis. Rather than simply discussing what should happen, participants are placed in a situation where they have to make decisions, coordinate with others, and respond as the scenario develops.

The concept is simple. Participants are placed in a realistic work environment and assigned roles that they would typically have in a crisis, such as IT, Communications, Management, Finance, HR, and Legal. They are not simply asked to discuss how they would respond to a cyberattack. They are placed in the middle of one. Inside the simulator, a series of coordinated cyberattacks unfolds in real time. New information appears, systems are disrupted, unexpected events occur, and the situation continues to evolve as participants respond. This makes it possible to observe what happens when people have to make decisions under pressure, often without having all the information they need. It also reveals their natural reflexes, how they deal with stress and uncertainty, how they communicate, and how well they work together when the pressure increases. In other words, ROOM#42 does not only test what people know. It shows how they actually behave when the crisis becomes real.

The learning process follows a continuous cycle. It starts by studying relevant threats and real-world incidents to identify the situations that participants may realistically face. These insights are then used to design and validate scenarios that reflect the complexity and uncertainty of a real crisis. Participants are observed as the scenario unfolds, not only to see what decisions they make, but also how they communicate, coordinate, and adapt as the situation changes. The exercise does not end when the scenario is over. The debrief is where much of the real learning takes place. Participants can review what happened, understand why certain decisions were made, identify weaknesses and agree on what needs to change. This final step is critical. An exercise that produces observations but does not lead to concrete improvements is little more than an interesting experience. The real value comes from turning what was learned into changes that make the next response better.

The purpose of a simulation such as ROOM#42 is not simply to test wheth-

er a company has a crisis plan. It is to assess its overall level of maturity and, more importantly, its ability to respond effectively when faced with a cyber crisis. The exercise provides a practical diagnosis of what works, where weaknesses exist, and which areas require improvement. These findings can then be used to define targeted training and concrete actions to strengthen the response. ROOM#42 is therefore not only an assessment tool but also a powerful training environment. Participants have the opportunity to put their knowledge and skills into practice, make decisions under pressure, and experience situations that would be difficult to reproduce in a classroom or through traditional training. They learn by doing, making mistakes, adapting and seeing the consequences of their decisions in a safe environment.

This practical experience is essential for turning knowledge into real capability. Understanding a procedure is one thing; being able to apply it when information is incomplete, time is limited, and pressure is high is something very different. By moving beyond theory and allowing people to practice in realistic situations, simulation helps develop the practical judgement, reflexes, and confidence needed to respond effectively when a real crisis occurs.

The ROOM#42 observations from 2024 and 2025 provide useful insight into some of the difficulties that can arise during a cyber crisis. These findings reflect the behaviour of participants within the specific scenarios and conditions of the ROOM#42 exercises and should not be treated as statistics representing the cybersecurity maturity of Luxembourg companies as a whole. Their value lies elsewhere: they show how people and teams can react when confronted with realistic situations, incomplete information, and time pressure.

The ransomware scenarios provide a good example of the importance of time and preparedness. The observations show that 41% of participants took more than fifteen minutes to detect and react to a crypto-ransomware scenario, while 69% of the countermeasures implemented were considered insufficient. These figures highlight a broader challenge: responding to ransomware is not a single technical action, but a combination of detection, decision-making, containment, business continuity and recovery. Technology can help detect and contain the threat. However, people still need to decide what to do, how quickly to act, and how to balance containing the attack with keeping critical business activities running. In a ransomware crisis, the quality and speed of these decisions can be just as important as the technology used to defend against the attack.

The observations concerning fake news reveal another important dimension of modern cyber crisis management. A cyberattack may target not only systems and data, but also information and trust. In the exercise scenarios, 81% of participants did not apply the correct reflexes when confronted with fake news. This is particularly significant in an environment where AI can make fabricated messages, images, voices and other forms of content increasingly

convincing and difficult to distinguish from genuine information. This creates a new challenge for crisis management. While security teams are trying to understand and contain the technical attack, false information may be spreading at the same time. Employees, customers, partners and the public may all be receiving different versions of what is happening. A future cyber crisis may therefore involve two battles at once: one against the attack on the technology, and another against the manipulation of information and trust.

The defacement scenario provides another useful lesson. The observation that 81% of participants did not know the correct first steps illustrates the difference between awareness and operational readiness. A company may know that website defacement is possible and may even have a documented procedure, but this does not mean that people will immediately know what to do when it actually happens. Under pressure, the first actions can make a significant difference. Evidence may need to be preserved, the affected service may need to be isolated, and communication needs to be coordinated. People need to know who should take the lead, what should be done first, and what actions could make the situation worse. These responses become much more natural when they have been practiced beforehand. This is one of the key benefits of simulation: it turns procedures that exist on paper into practical reflexes that people can rely on when a real crisis occurs.

Observations from ROOM#42 sessions point to a clear conclusion: teams are better equipped to limit the impact of a cyber crisis when they practice regularly, understand their responsibilities, and follow a simple, structured approach to managing incidents as they unfold. Technical expertise matters, but it must be paired with clear priorities, coordinated action, and the ability to adapt as the situation changes. The key principles of effective crisis management observed during ROOM#42 can be summarised as follows:

- **Detecting incidents and activating the crisis response**: Recognising when an incident requires escalation, understanding its potential impact, and activating the appropriate crisis management procedures without unnecessary delay.
- **Establishing a clear direction**: Building a shared understanding of the situation and defining how the response should proceed, including which essential business activities must continue and what minimum level of operations needs to be maintained.
- **Setting clear priorities**: Identifying the systems, services, data, and business functions that require immediate protection or recovery, and ensuring that everyone understands what matters most.
- **Maintaining a crisis log**: Recording key events, decisions and actions, including their dates and times, to maintain a reliable timeline and support coordination, accountability and later review.

- **Assessing the impact:** Understanding the technical and business consequences of the incidents, evaluating the risks, and using this information to determine which issues require attention first.
- **Developing and maintaining an action plan:** Translating the impact assessment into a clear, practical plan that defines the actions to be taken, the people responsible, and the priorities to be followed. The plan must also be updated as new information becomes available.
- **Maintaining team resilience:** Ensuring that people can work together, manage pressure, adapt to changing circumstances and sustain their efforts throughout a crisis, including when the response extends over several hours or days.
- **The ability to determine when a crisis is over:** Establishing clear criteria to assess when the crisis can officially be declared over and when normal operations can safely resume.

Taken together, these principles provide a practical framework for managing a cyber crisis. Their effectiveness depends not on having a perfect plan, but on whether people understand the framework, practice it regularly, and can apply it when time is limited and the situation remains uncertain.

Scan the QR code for a preview of the ROOM#42 exercise.



4. Managing the Crisis: Decisions, Coordination and Communication

The move from preparation to crisis management begins when action is needed. In reality, this transition is rarely clear-cut. A suspicious event may first look like a normal technical incident, but as more information emerges, it may become clear that the situation is much more serious. At that point, many questions may remain unanswered: Is this a malicious act or simply a system failure? If it is malicious, is the attacker still present? How far has the compromise spread? What information may have been accessed? Yet decisions may already need to be made.

One of the first decisions is whether to escalate the situation. Escalation does not mean declaring a catastrophe. It means recognizing that the potential impact is serious enough to involve additional people and resources. When the

potential impact is high and there is still a lot of uncertainty, it is often safer to escalate early rather than wait for all the facts. The cost of involving the crisis team unnecessarily is usually much lower than discovering hours later that a serious incident should have been escalated from the start.

Having worked in the banking sector earlier in my career, this part brings back many memories. When an incident affected a core banking application or the cardholder environment, clear escalation procedures were in place. This could mean calling senior management in the middle of the night and bringing everyone into a Webex call that sometimes lasted until the afternoon. At the time, it could feel excessive, but the principle was simple: when the potential impact is high, the right people need to be involved early. You do not wait for complete certainty before starting to manage a potential crisis.

Once the crisis response begins, the priority is to establish control. This does not mean solving the technical problem immediately. It means creating enough structure to avoid confusion and uncontrolled reactions. Everyone should understand what is happening, who is coordinating the response, what information is available, and which decisions need immediate attention. A timeline should also be created from the start. As a crisis develops, memories become less reliable, people may remember events differently, and decisions may later need to be reviewed. Keeping a clear record of what happened, when it happened, and what actions were taken makes it much easier to understand how the situation developed and why particular decisions were made.

The first hour is not about having all the answers. It is about getting everyone to understand what is happening. The crisis team needs to work out what happened, when it started, which systems may be affected, whether the incident or the attacker is still active, which business services are being disrupted, and what is still unclear. It is important to be honest about what is known and what is not. Trying to appear certain when the facts are still unclear can create a false sense of confidence and lead to the wrong decisions. Being open about uncertainty makes it easier to take precautionary action and change course as more information becomes available.

Impact assessment should not stop at the technical side of the incident. Knowing that three servers have been compromised does not tell management what this actually means for the business. The important question is what those servers are used for. If they support a non-critical internal application, the impact may be limited. If they support customer transactions or an important business process, the situation can be much more serious. Technical information therefore needs to be translated into business impact.

This is also why having people from different areas involved in the crisis is so important. The security team understands what is happening technically, while the business understands what the affected systems mean in practice.

Legal specialists can identify potential obligations, communications teams can consider what needs to be shared with customers and other stakeholders, and senior management can look at the wider consequences for the business. Bringing these different perspectives together gives the crisis team a much clearer picture of the situation and helps it make better decisions.

Making decisions with incomplete information is one of the hardest parts of crisis management. Leaders cannot always wait until they know exactly what happened. They need to identify which decisions need to be made immediately and which ones can wait. It is also important to be clear about the assumptions behind each decision. For example, if there is a possibility that an attacker still has access to the environment, the team may decide to take a more cautious approach to containment. If new information later shows that the attacker is no longer present, the approach can be changed. What matters is that the team understands why the decision was made and what information it was based on.

Containment is a good example of this challenge. The natural reaction to a cyberattack is often to isolate affected systems as quickly as possible. Sometimes that is exactly what needs to happen. But shutting down or disconnecting a system can also disrupt critical business services. A compromised system may therefore need to remain operational for a short period while a safer way to contain the threat is found. There is no single answer that works in every situation. The right decision depends on the potential impact, the level of risk, and what the business needs to keep operating.

This is where cybersecurity and business continuity come together. The technical team may want to stop the attack as quickly as possible, while the business needs to keep essential services running. Neither side can simply be ignored. The crisis team needs to understand the trade-off and make a clear decision based on the risks and the potential impact on the business.

The people involved in the response also need to be considered. A cyber crisis can last for many hours or even several days, and fatigue can quickly affect concentration and decision-making. Someone who is highly effective at the beginning of an incident may make mistakes after working for too long without a break. Crisis preparedness therefore needs to take human endurance into account. Shifts, backup personnel, food, rest and other basic needs may seem secondary during a crisis, but they are essential to keeping people effective and the response going.

External support should be brought into the response when needed. A company may have strong internal cybersecurity capabilities and still benefit from specialists who have dealt with similar incidents before. They can also challenge assumptions that the internal team may have started to take for granted, which can be extremely valuable during a crisis. The technical response, however, is only one part of the crisis. Communication needs to happen at the same time

and can quickly become a crisis of its own. Employees will want to know what is happening. Customers may want to know whether their information is safe. Regulators may require information, suppliers may need instructions, and journalists may start asking questions. Managing these different audiences while the facts are still developing is therefore another important part of the crisis response.

Internal communication is especially important during a crisis. Employees are part of the response, whether they are formally involved or not. If they do not receive clear information, they will naturally try to make sense of what they see. Rumors can spread, people may continue using systems that have been compromised, and employees may give customers different or even contradictory information.

The crisis structure should therefore make it clear who communicates externally, who is responsible for interacting with the attacker, and how information is shared internally. The goal is not to remove the human side of the response, but to make sure that everyone knows their role and that decisions are coordinated. Otherwise, the attacker may take advantage of confusion, inconsistent messages, or gaps in decision-making.

A well-managed cyber crisis is not one where everything is known from the beginning. It is one where people can continue to act effectively even when many things remain uncertain. The crisis team builds a shared understanding of what is happening, makes decisions based on the best information available, communicates clearly, protects the most important business functions, and adapts its response as new information emerges. This is what turns a series of individual actions into a coordinated response.

5. Cyber Crisis Management in the Age of AI

Artificial intelligence is changing the way cyber crises unfold. Its role goes beyond that of another cybersecurity tool, such as a monitoring or threat detection system. AI can process information faster, generate content in seconds, and help both attackers and defenders carry out certain tasks more quickly. As a result, the pace of a cyber crisis can change, leaving less time to understand what is happening, make decisions, and respond effectively. For defenders, AI offers some clear advantages. Modern security systems generate huge amounts of information, including logs, alerts, endpoint events, network activity, and threat intelligence. Going through all this information manually can take a lot of time, and important signs of an attack may be missed. AI can help by identifying patterns, linking related events, and summarising the information that matters most. Instead of spending hours trying to connect individual alerts, a security analyst may be able to build a clearer picture of a potential attack within minutes.

This can be particularly useful in the early stages of a crisis, when the security team has to deal with hundreds or even thousands of alerts and signals, often with very little time to understand what is happening. AI can help make sense of this information by bringing related events together, highlighting patterns and identifying which issues need attention first. This gives analysts a clearer picture of the situation and helps them focus on the events that matter most.

But AI does not eliminate uncertainty. Its assessments can be wrong, incomplete or based on a misunderstanding of the situation. AI can help the crisis team make sense of what is happening, but its conclusions should never be accepted without question. This is why a human must remain involved in the decision-making process. AI can support people by providing information and highlighting potential risks, but people must be able to question its conclusions, check the facts and decide what action to take. The same applies to decision-making. AI can recommend actions, predict possible consequences and automatically carry out certain predefined responses. This can be very useful when every second counts. For example, if an account is confirmed to be compromised, automatically restricting its access can help prevent further damage. However, not every decision should be left to automation. Disconnecting a critical production system, shutting down an essential service or sharing sensitive information externally can have consequences that go far beyond cybersecurity. These decisions require an understanding of the wider business impact, which is why human judgement remains essential.

Companies therefore need to be clear about where AI can assist and where people must remain in control. Some actions can be safely automated because their impact is limited and the conditions for taking them are well understood. Other actions should require human approval because they could have significant consequences for the business. The most important decisions should always have a person accountable for them, regardless of how advanced the AI system becomes.

AI is also making cyberattacks more sophisticated and harder to detect. Phishing emails, for example, no longer need to contain spelling mistakes or awkward wording that immediately raises suspicion. Attackers can use AI to create convincing messages, adapt them to specific targets, produce realistic documents, and even imitate voices. Imagine a senior employee receiving a voice message that sounds exactly like the CEO. The message requests an urgent payment and explains that the CEO is unavailable for further discussion. The request fits an ongoing business situation, and the employee is already under pressure because the company is dealing with a cyber incident. Nothing about the message seems obviously suspicious. In this situation, cybersecurity training alone may not be enough. Even experienced employees can be misled by convincing imitations, especially when they are under pressure. Companies

therefore need clear procedures for independently verifying unusual or sensitive requests before taking action, regardless of how genuine they appear.

As AI becomes better at imitating people and making fraudulent requests look genuine, companies need stronger ways to verify that those requests are legitimate. The answer is not to tell employees to distrust everything they receive. That would make everyday business operations unnecessarily difficult. Instead, companies need clear verification procedures for high-risk actions, such as urgent payments, changes to access rights, or requests for sensitive information. These procedures should be followed even when a request appears to come from someone they know and trust. AI can also help once the immediate crisis is under control. After an incident, the team may have large amounts of information to review, including logs, forensic reports, messages, and records of decisions taken during the response. AI can help summarise these findings, identify patterns and suggest possible explanations for further investigation. It may reveal connections between incidents that initially appeared unrelated, such as several attacks exploiting the same weakness or recurring communication problems that contributed to failures during the response. These insights can help the team understand not only what happened, but also what needs to change to prevent similar problems in the future.

After the crisis, it is also important to look at how AI was used during the response. Did it help the team respond faster and make better decisions, or did it create new problems? Did an AI system provide incorrect or misleading information? Did employees use AI tools that had not been approved? Was sensitive information shared with an AI service? Did the team rely too heavily on AI-generated recommendations without checking them? These questions should be part of the lessons-learned process. The purpose is not to decide whether AI was simply good or bad, but to understand how it affected the response. Where did it genuinely help? Where did it create additional risks? What should be changed before the next crisis? Answering these questions helps companies understand where AI can add value, where human oversight is needed, and what safeguards should be put in place before using it again.

The most important question therefore is not whether AI should be used during a crisis, but how it should be governed. Companies need to be clear about which decisions AI can make on its own, which require human approval, and which should always remain under human control. They also need to make sure that decisions supported by AI can be understood, reviewed, and reconstructed afterwards. This is important not only for accountability, but also for learning from the crisis and improving future responses. This becomes even more important when the attacker is using AI as well. Defenders may be facing an adversary that can generate convincing content in seconds, automate reconnaissance, adapt social-engineering campaigns, and analyse large amounts of

information with very little human effort. The speed and scale of these activities can change the nature of the attack itself. Defenders therefore cannot assume that traditional response times will always be fast enough. They need to understand where AI can help them respond faster, while making sure that speed does not come at the expense of control and good judgement.

AI creates a race against time, but speed alone is not the objective. Companies need to become faster without becoming reckless. Automation should help accelerate the right decisions, not simply make every decision faster. This is why the three themes of Time, People, and Skills remain just as relevant in an increasingly automated environment. Time becomes even more important because AI can accelerate both attacks and responses. People remain essential because someone must set priorities, challenge assumptions and take responsibility for the decisions that are made. Skills become increasingly important because employees need to understand not only cybersecurity, but also how AI can change the speed, reliability and nature of the information they receive. The more AI becomes part of the response, the more important it becomes to know when to trust it, when to question it, and when to rely on human judgement.

The future is unlikely to be a choice between human intelligence and artificial intelligence. A more realistic model is one in which humans and machines work together, each contributing what the other does best. Machines can process vast amounts of information at a speed that humans cannot match. People can provide context, judgement, responsibility, and ethical direction. The real challenge is to design this relationship carefully, so that the strengths of one help compensate for the limitations of the other. Companies that succeed in this environment will not necessarily be those with the most advanced AI systems. They will be the ones that understand where AI can create value, where it introduces new risks, and where human judgement remains essential. Ultimately, the goal is not to replace people with AI, but to use AI in a way that helps people make better decisions and respond more effectively when it matters most.

6. Beyond the Crisis: Recovery, Lessons Learned and Resilience

When a cyber crisis is brought under control, the natural reaction is to get things back to normal as quickly as possible. Systems need to be restored, employees need to resume their work, customers expect services to be available again, and management wants to move forward. After hours or even days of stress and uncertainty, it is understandable that everyone wants to put the incident behind them. However, getting back to normal does not mean that everything is over. There is still work to be done, and how the company handles this period can make a real difference to its recovery and its ability to deal with future incidents.

One of the first questions to answer is when the crisis can officially be considered over. The fact that the immediate incident appears to have been resolved or that the affected systems are working again does not necessarily mean the situation is fully under control. Whether the crisis was caused by a cyberattack, a system malfunction, or another type of disruption, the company needs to make sure that the underlying problem has been addressed, that essential business activities can safely resume, and that any remaining risks are understood. In some cases, the immediate issue may be resolved while important weaknesses or uncertainties remain. The team must decide whether these can be managed as part of normal operations or whether crisis-level coordination is still necessary. To avoid confusion, the criteria for declaring the crisis over should be agreed in advance, and the final decision should be made by people with the appropriate authority and a clear understanding of the situation.

Once the situation is under control, the focus shifts to restoring normal operations. This involves more than simply switching systems back on or recovering data from backups. Before systems are put back into production, the team needs to make sure they are safe to use and that the causes of the disruption have been addressed. Depending on the incident, this may mean securing compromised accounts, reviewing access rights, fixing vulnerabilities, or checking the integrity of restored data. Rushing this process can create new problems, potentially leading to another disruption or, in the case of a cyberattack, allowing the attacker to regain access. Recovery therefore requires a balance between speed and caution. The goal is not just to get systems running again, but to restore them safely and ensure that essential business activities can resume reliably. Recovery requires clear priorities. Not every system needs to be restored at once, and trying to recover everything simultaneously can put unnecessary pressure on teams. The company must identify its most critical services, understand their dependencies, and determine what needs to be restored first to keep essential business activities running. These priorities should be based on business impact, not just technical importance. IT and business teams must work together to decide what needs immediate attention and what can safely wait.

Once the situation is stable, the company also needs to understand what happened. The team should investigate how the incident started, how the attacker gained access if it was a cyberattack, which systems and data were affected, and how the situation developed. Logs, system records, forensic findings and the crisis timeline can help reconstruct the events and explain the decisions made during the response. Not every question will necessarily have a clear answer, but conclusions should be based on available evidence, with any remaining uncertainties clearly identified. The company also needs to assess how it managed the crisis. Was the incident detected and escalated quickly enough? Did the team have the information it needed to make decisions? Were

responsibilities clear, and did the different teams work well together? Were employees, customers and other stakeholders kept informed? Answering these questions helps identify weaknesses in procedures, skills, communication, and decision-making. Often, several factors contribute to the outcome. Looking at them together helps the company understand what went wrong and what needs to change to handle future incidents more effectively.

The real value of a post-crisis review lies in turning lessons learned into concrete improvements. If escalation was too slow, the criteria and procedures need to be clarified and tested. If communication failed because the usual tools were unavailable, alternative methods need to be reviewed. If recovery took too long, backups and restoration procedures may need to be improved. If people were unsure who could make critical decisions, responsibilities and authority must be clarified. Each action should have a clear owner and deadline, with progress monitored to ensure that the improvements are actually implemented. Otherwise, the same weaknesses may remain until the next crisis exposes them again.

The end of the crisis does not mean the company's obligations are over. Depending on the incident, regulatory reporting, customer communications, contractual commitments, insurance claims and legal enquiries may continue long after operations have returned to normal. The company must keep accurate records, coordinate communications and ensure that someone remains responsible for following up on these matters. It should also be ready to explain what happened, what was affected, how it responded and what has been done to prevent a similar incident from happening again.

If AI was used during the response, the company should also review how it was used. Did it help the team understand the incident faster? Did it provide inaccurate information or inappropriate recommendations? Were automated actions effective? Did employees use unapproved tools or share sensitive information with external AI services? The aim is to understand where AI helped, where it created risks, and what safeguards are needed. As AI becomes more common in cybersecurity and crisis management, reviewing its use will be an important part of learning from incidents and improving future responses.

A cyber crisis should not be considered a success simply because systems are back online and business operations have resumed. Recovery is only one part of the process. The real test is what happens afterwards: whether the company understands what went wrong, learns from the decisions made under pressure, and turns those lessons into meaningful improvements. No company can predict every future attack, but it can use each incident to improve its ability to respond, recover, and adapt. Ultimately, the purpose of looking beyond the crisis is not just to return to normal, but to be better prepared for what comes next.

7. Resilience is an Ongoing Process…

No company can guarantee that it will never face a cyber crisis. But it can prepare to respond, protect what matters most, and recover when things go wrong. As this chapter has shown, effective crisis management is not just about technology or written procedures. It depends on people knowing what to do, making decisions under pressure and working together when uncertainty is greatest.

AI is changing the speed and nature of cyber crises, creating new opportunities for defense as well as new risks. Yet even with more advanced technology, human judgment, clear responsibilities, and effective communication remain essential. The real challenge is not simply responding faster but making the right decisions at the right time. Ultimately, resilience is built before the crisis through preparation, practice, and continuous learning. It comes back to three fundamental elements: Time, People, and Skills. How we use available time, how we work together, and how well we prepare will determine our ability to face a crisis and emerge stronger.

References

ENISA Threat Landscape 2026
https://www.enisa.europa.eu/topics/cyber-threats/threat-landscape
Best Practices for Cyber Crisis Management published by ENISA https://www.enisa.europa.eu/publications/best-practices-for-cyber-crisis-management
Incident Response Recommendations and Considerations for Cybersecurity Risk Management (SP 800-61 Rev. 3) published by the NIST.
https://doi.org/10.6028/NIST.SP.800-61r3
Cyber Blueprint for Cybersecurity Crisis Management (Council Recommendation (EU) 2025/3445) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32025H03445
The ENISA Cybersecurity Exercise Methodology
https://www.enisa.europa.eu/publications/the-enisa-cybersecurity-exercise-methodology
Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile https://doi.org/10.6028/NIST.AI.600-1
Les Fondamentaux de la Gestion de Crise Cyber by Laurane Raimodo
ROOM#42 The Cyberattack Simulation made in Luxembourg by the National Cybersecurity Competence Center (NC3)
https://room42.lu/

SINGAPORE'S NATIONAL CYBERSECURITY STRATEGY AND RESILIENCE MODEL

Danielle Yeow
National University of Singapore, Singapore

Abstract
This chapter examines Singapore's national cybersecurity strategy and resilience model, highlighting its legally grounded, risk-based and framework-driven approach to securing an open, highly connected digital economy. It explores the strategic drivers of Singapore's cybersecurity posture, including its role as a global financial and digital infrastructure hub, growing cyber threats and the security challenges posed by artificial intelligence. The chapter analyses the institutional and regulatory architecture underpinning national cyber resilience, with particular attention to the Cybersecurity Act 2018 and its 2024 amendments, critical information infrastructure protection, supply chain security and emerging technologies. It also considers the relationship between digital sovereignty, cross-border data flows and public-private partnerships. Through an examination of operational technology, agentic AI, quantum computing, election integrity and disinformation, the chapter illustrates how Singapore combines binding regulation with adaptive guidance, codes of practice and certification mechanisms. Finally, it assesses Singapore's contribution to ASEAN cybersecurity cooperation and regional norm-building, presenting its model as a relevant reference for small states seeking to balance national security, digital connectivity, innovation and resilience amid intensifying geopolitical competition.

1. Introduction: An Open and Connected Economy

Singapore's growth strategy rests on an open and connected economy (Ministry of Trade and Industry[1], 2017; Ministry of Trade and Industry, 2025). The digital economy contributed about S$128.1 billion (approximately US$100.4 billion) or 18.6% of Singapore's gross domestic product in 2024 (Infocomm Media Development Authority, 2025). Singapore is a leading financial hub (Global Financial Centres Index, Z/Yen Group and China Development Institute, 2026) and a trusted regional hub for digital infrastructure, hosting submarine cables linking Singapore and the region to the world as well as more than 1.6 gigawatts of data centre capacity (Ministry of Digital Development and Information, 2026a). Public services are highly digitized, and critical infrastructure depends heavily on networked systems. Singapore also aims to be a global leader in artificial intelligence

[1] Renamed to the Ministry of Energy, Trade and Industry wef 1 October 2026.

(AI) solutions and an AI-empowered economy; as well as to strengthen Singapore's role as a connected and trusted hub (Economic Strategy Review Committees, 2026). The global information and communications technology (ICT) supply chains and interconnected digital ecosystems create intertwined digital dependencies that a determined adversary can exploit, which requires the effective management of cybersecurity and resilience risks.

That risk is not hypothetical. On 18 July 2025, the Coordinating Minister for National Security disclosed that an advanced persistent threat (APT) actor, UNC3886, had attacked Singapore's critical information infrastructure (CII) and emphasized the threat to national security posed by APTs and foreign actors (Cyber Security Agency of Singapore, 2025a). The Cyber Security Agency of Singapore (CSA) and the Infocomm Media Development Authority (IMDA) subsequently revealed that all four of Singapore's major telecommunications operators had been targeted by UNC3886. Over a period of over eleven months, more than 100 defenders across several agencies took part in a coordinated whole of Government response, codenamed Operation CYBER GUARDIAN, in partnership with the telecommunications operators, to contain the breach (CSA, 2026h). Unlike previous cyber intrusions, UNC3886 posed a more serious threat as it targeted systems that directly deliver essential services to the public. Had it proceeded further, telecommunications or internet access could have been seriously disrupted, with downstream effects on other essential services like banking and finance, transport, and medical services. Successful attacks can also undermine trust and confidence in Singapore and Singapore's economic security (Teo, 2026a).

AI, while central to Singapore's economic growth strategy (Ministry of Trade and Industry, 2026b), also compounds cybersecurity risk. CSA noted that frontier AI models can reportedly shorten the time needed to find vulnerabilities and engineer exploits from months to hours (CSA, 2026a); and warned that agentic AI could automate large parts of the cyber kill chain, lower the entry barrier for less sophisticated threat actors, and compress attacks that once unfolded over days into hours (CSA, 2026f).

This chapter argues that Singapore's response is best understood as a legally grounded, framework-driven model whose risk posture is continually recalibrated. Singapore asserts regulatory authority where systemic risk demands it and restricts data localisation that erodes the functioning of the international digital economy and Singapore's position as a connected hub. Hard law is reserved for systemically important infrastructure, and is accompanied by evolving guidelines, codes and certifications that shape the wider ecosystem. The approach is marked by public-private partnerships that address cybersecurity risks. This model also finds its way into Singapore's approach towards regional and global norm-building.

The chapter moves from strategy to the institutional and legal framework, before examining four thematic challenges, and finally to Singapore's regional normative role.

2. Cybersecurity and Resilience Strategy

Cybersecurity is a necessity and key enabler for Singapore's digital economy (CSA, 2021a). Singapore's approach is best encapsulated by Singapore's then Deputy Prime Minister Teo Chee Hean's speech on the opportunities of the digital economy (CSA, 2018a):

1. Digital technologies and Industry 4.0 are examples of where resilience and innovation come together. The rapid growth of digital technologies, such as the Internet of Things, cloud computing and Artificial Intelligence, has broken down walls and opened up new opportunities for everyone.
2. Innovation is key in an increasingly digital future; but so too is resilience provided by robust cybersecurity. Cybersecurity is the key enabler for us to confidently seize the new opportunities from the digital revolution, and foster trust in a digital future. Conversely, weak cybersecurity leaves us exposed to new and disruptive cyber threats.
3. To build a resilient and innovative digital community together, we need to invest in three elements: Cyber Defence, Cyber Innovation and Cyber Partnerships.

Strategic Pillars

Singapore's Cybersecurity Strategy 2021 comprises three strategic pillars: building resilient infrastructure, enabling a safer cyberspace, and enhancing international cyber cooperation; as well as two foundational enablers: a vibrant cybersecurity ecosystem and a robust cyber talent pipeline (CSA, 2021b). The first pillar places CIIs at the core while extending attention to government systems and important entities and systems beyond CIIs. The second pillar addresses securing digital infrastructure, devices and applications; safeguarding cyberspace activities; and empowering a cyber-savvy population. The third is focused on advancing the development and implementation of voluntary, non-binding norms alongside international law; capacity building and developing technical and interoperable cybersecurity standards; and international efforts to combat cross-border cyber threats.

While regulation remains a key lever, the 2021 Strategy observes that a purely regulatory approach could "engender a compliance mindset which is not desirable in a fast-evolving environment" and exhorted enterprises and organisations

to adopt a risk management mindset and frameworks (CSA, 2021b: 9). In tandem with the evolution of the cyber threat landscape, there has been a recalibration towards regulation. Examples include the 2024 amendments to the Cybersecurity Act 2018, mandatory certification requirements, and a new Code of Practice for CIIs (discussed below). This is likely due to the significant increase in infected infrastructure cases detected in Singapore in 2025, numbering 284,300 and reflecting a 142% increase from 2024. The increase was driven primarily by persistent malicious infrastructure activities, facilitated by malware-as-a-service offerings and poorly secured consumer Internet-of-Things devices (CSA, 2026f). At the same time, a flexible phased approach is maintained for rapidly evolving technological developments like frontier AI. There are currently plans to review Singapore's national strategy and masterplan in 2026 in response to the disruptive wave of AI-enabled cyber threats (CSA, 2026g: 43).

Digital Sovereignty

Singapore's cybersecurity strategy and resilience model is also influenced by its approach towards digital sovereignty and data localisation. Strategic autonomy in the sense of self-sufficiency is unrealistic, given Singapore's heavy reliance on foreign cloud services and hardware. On the other hand, territorial control over data and networks would undermine the very connectivity on which the hub economy depends. Singapore has consistently favoured trusted cross-border data flows and resisted unjustified data localisation (Teo, 2022), calling for more data connectivity and less data localisation on the international front. The premise being that cross-border data flows are essential to the international digital economy. At the same time, Singapore acknowledges legitimate concerns about national security. The focus for policy-makers, working with the technology community, is on enhancing data connectivity while taking into account issues of data sovereignty, and ensuring common data standards across countries to enable the free flow of data in an environment of trust and security (Menon, 2018).

The Cybersecurity Act 2018, as amended, reflects this deliberative approach (Singapore Statutes Online, 2025). The CA regulates by function and control rather than solely by location. Virtual systems, third-party-owned systems and, in defined circumstances, systems located wholly overseas may be designated where they are necessary to the delivery of an essential service in Singapore (Singapore Parliament, 2024). Additionally, the recently tabled Digital Infrastructure Bill (Singapore Statutes Online, 2026a) establishes new location-agnostic licencing regimes to strengthen the security and resilience of major cloud services (that generate revenue above a specified threshold from Singapore users) and data centres (with capacity above a specific threshold).

On the international front, Singapore promotes cross-border digital trade through digital economy agreements (DEAs) (Ministry of Trade and Industry,

2026a). The DEAs seek to align digital rules and technical standards across jurisdictions, facilitate interoperability between digital systems, and support trusted and secure cross-border data flows to safeguard personal data.

In short, Singapore decides which functions are systemically important and insists on their assured accountability, while enabling trusted and secure free cross-border data flow.

3. Institutional Architecture and Governance Framework

Network of Regimes and Instruments

Singapore's cybersecurity governance landscape reflects a network of regimes and instruments comprising a mix of mandatory regulations, codes and guidance that address different aspects of cybersecurity and digital resilience across different sectors.

The CSA was established in 2015 to protect Singapore's cyberspace, enhance cybersecurity awareness, and build national resilience against cyber threats. CSA sits within the Prime Minister's Office and is managed by the Ministry of Digital Development and Information (MDDI), which provides central strategic direction, while sectoral regulators oversee their respective domains.

For instance, the Monetary Authority of Singapore (MAS) issued Technology Risk Management Guidelines (TRM Guidelines) that are broadly applicable across Singapore's financial services sectors and provide guidance on the management of technology risk. The TRM Guidelines contain risk management principles and industry best practices to establish robust technology risk management frameworks and maintain cyber resilience of the financial institutions' (FIs) infocommunications technology (IT) infrastructure. They are not in themselves legally binding, but observance is considered as part of the regulator's risk assessment of the FIs (Wong Partnership, 2021).

The TRM Guidelines were substantially revised in 2021 to include new guidance on effective cyber surveillance, secure software development, adversarial attack simulation exercises, and management of cyber risks posed by emerging technologies (Monetary Authority of Singapore, 2021). They operate on a principles-based approach, setting out expectations while allowing FIs flexibility in the implementation of internal controls. Notices and circulars by the MAS complement the TRM Guidelines, establishing legally binding requirements such as the notification of IT security incidents and implementation of IT controls to protect customer information from unauthorised access or disclosure (Monetary Authority of Singapore, 2024). In light of the increasing digitalisation and evolving risk landscape, a consultation paper issued in June 2026 proposed amendments to the TRM Guidelines. The amendments sought to fur-

ther strengthen FIs' technology resilience, covering key areas such as IT asset management, IT risk assessment and monitoring, change management controls, and data backup and recovery (Monetary Authority of Singapore, 2026).

Another sectoral lead is IMDA, which is responsible for both developing and regulating the infocomm and media sectors. IMDA's governance framework includes a mandatory Telecommunication Cybersecurity Code of Practice on cyber security preparedness of major Internet Service Providers (ISPs), the establishment of the Infocommunications Singapore Computer Emergency Response Team (CERT), and a Cyber Security Vulnerability Reporting Guide (IMDA, 2026b).

Personal data protection falls under the purview of the Personal Data Protection Commission (PDPC), which is responsible for administering the Personal Data Protection Act 2012 (PDPA). Section 24 of the PDPA is the main cybersecurity lever. It imposes an obligation on organisations to "protect personal data in its possession or under its control by making reasonable security arrangements to prevent" the unauthorised access, collection, use, disclosure, modification or disposal and so on, and the loss of any storage medium holding personal data (Singapore Statutes Online, 2026c). A later amendment added an obligation to notify the regulator of data breaches (Singapore Statutes Online, 2026c: 26A-26E). A number of regulatory guidelines, including advisory and sector-specific guidelines, have also been issued to support organisations' implementation and compliance with the PDPA while promoting responsible data handling practices (PDPC, n.d.).

In relation to key healthcare information stored on a national electronic health record database, the new Health information Act 2026 requires healthcare providers, retail pharmacies, and health information management system vendors to implement reasonable security measures to safeguard health data (Singapore Statutes Online, 2026b). These measures include strict access controls; cybersecurity and data security requirements; audit logs; as well as mandatory reporting of significant data breaches and cybersecurity incidents to the Ministry of Health.

Cybersecurity Act 2018 and its 2024 Amendments

The Cybersecurity Act 2018 (CA) establishes the primary statutory framework for the oversight and maintenance of national cybersecurity in Singapore (Singapore Statutes Online, 2026a). The CA:

(a) establishes a CII designation framework and obligations of CII owners to proactively protect the CII from cyber-attacks. The relevant CII sectors are energy, water, banking and finance, healthcare, transport, infocomm, media, security and emergency services and Government;
(b) empowers CSA to investigate cybersecurity threats and incidents and

prevent further harm or cybersecurity incidents from arising. Such powers are calibrated according to the severity of the threat or incident and required response measures;

(c) facilitates the timely sharing of cybersecurity information as to system vulnerabilities for the effective prevention of cyber incidents;

(d) establishes a licensing framework for cybersecurity service providers. The framework reflects a balance between security needs and the development of a vibrant cybersecurity ecosystem. In a light touch approach, this is limited to the licensing of penetration testers and managed security operations monitoring centres (prioritised due to their access to sensitive client information and their significant impact on the overall security landscape in Singapore).

The Cybersecurity (Amendment) Act 2024 was adopted on 23 May 2024, with some provisions taking effect on 31 October 2025 (Singapore Statutes Online, 2024a). The amendments allow the CA to keep pace with the developments in the cyber threat landscape and Singapore's evolving technological operating context (Singapore Parliament, 2024). The amendments:

(a) authorise the designation of virtual systems and third party owned systems, including overseas systems, that are relied on in the provision of essential services as CII. This ensures that CII owners remain responsible for their cybersecurity and cyber resilience, even as they adopt new technological and business models such as the use of cloud computing. Cybersecurity incident reporting obligations are also expanded to include suspected APTs and incidents that occur in their supply chain;

(b) expands CSA's oversight to include new classes of regulated entities, termed Systems of Temporary Cybersecurity Concern (STCCs), Entities of Special Cybersecurity Interest (ESCI) and Foundational Digital Infrastructure (FDI). STCCs refers to systems facing heightened risk due to temporary events or situations, which could be detrimental to Singapore's national interests such as those supporting elections. The owners of such systems, whether located wholly or partly in Singapore, are required to undertake cybersecurity measures and report cybersecurity incidents affecting the STCC during a specified time limited period. ESCIs may be designated if they hold sensitive information or perform a function of national interest. Companies that provide foundational digital infrastructure services (such as cloud and data centre operators) will be regulated as FDI and required to adhere to cybersecurity codes and standards of practices, as well as cybersecurity incidents reporting requirements (CSA, 2025c).[2]

[2] As at time of writing, Parts 3C and 3D of the Cybersecurity Act 2018 (entities of special

The higher standard and stringent obligations imposed on CII owners reflects a calibrated approach that balances national security needs and business costs (Tan, 2026).

Cybersecurity Code of Practice (CCoP) for Critical Information Infrastructure

CII owners, designated providers of third-party CII, and other regulated entities and providers under the CA are required to comply with codes of practice or standards of performance issued by the Commissioner of Cybersecurity (CoC) from time to time (Singapore Statutes Online, 2026a: 35A). The CCoP for Critical Information Infrastructure (updated version issued on 29 July 2026) sets out the minimum requirements that CII owners are required to implement to ensure the cybersecurity of their CII and CII interconnected systems. However, CII owners are expected to implement measures beyond those set out in the CCoP (CSA, 2026d).[3]

The CCoP address a broad range of topics including the audit, governance, protection, detection, response and recovery, and cyber resilience requirements; cybersecurity training and awareness; operational technology security requirements; and the securing of interconnected systems that are owned, operated, or controlled by the CII owner. Additionally, the 2026 updates respond to the increasing APTs and AI-enabled threats and are intended to raise the cybersecurity posture of CII owners in alignment with the 2024 CA amendments (CSA, 2026e).

The new requirements relate to:

(a) strengthening board level accountability and senior management responsibility for cybersecurity;
(b) attaining Cyber Trust Mark Level 5 certification (CSA, 2026b);
(c) developing a cybersecurity exercise plan;
(d) facilitating the deployment of CSA-supported threat detection systems; and
(e) securing CII interconnected systems through measures including network segmentation, systems hardening, monitoring and detection etc (CSA, 2026e; Rajah & Tann Asia, 2026).

Additionally, a new CCoP for Cloud Service will be launched in the second half of 2026 to establish requirements for the secure deployment, operation and management of CII systems hosted on the cloud. CSP-companion guides

cybersecurity interest and major foundational digital infrastructure providers, respectively) are not yet in force.

[3] The updated CCoP took effect from date of issuance with most new obligations applying from 29 July 2027.

jointly developed by CSA and leading CSPs (Amazon Web Services, Google Cloud and Microsoft Azure) are also expected to be published alongside the new CCoP for Cloud Services (CSA, 2026e).

Certification and Labelling

As part of the broad efforts to better secure Singapore's cyberspace and improve cyber hygiene, CSA's Cybersecurity Certification Centre operates several certification and labelling schemes such as the Cybersecurity Labelling Scheme for IoT and Medical Devices.[4] The labelling schemes provide the security assurance that the product has been independently examined and tested, which ensures that it is securely designed, implemented, and appropriate in mitigating specified security threats.

CSA's Cyber Essentials certification equips organisations with security controls and best practices to address emerging cyber risks such as cloud security, OT security and AI security.[5] CSA's Cyber Trust certifications, which are suited for organisations with more extensive digitalised operations, help organisations establish good cyber security practices and measures via a risk-based approach.[6]

Digital Infrastructure Bill – Foundational Digital Infrastructure

The Digital Infrastructure Bill (DIB) was introduced in Parliament on 8 September 2026 (Singapore Statutes Online, 2026a). The Bill was developed in collaboration with a broad range industry stakeholders including data centre operators, clouds service providers (CSPs), and industry association and enterprise users. It complements the cybersecurity requirements for major foundational digital infrastructure set out in the CA and covers a broader range of operational resilience risks which can also disrupt the delivery of digital services. The DIB serves to strengthen the foundational infrastructure of Singapore's digital economy and leading regional AI hub ambitions. In essence, the Bill proposes the establishment of a licensing regime for major co-location and cloud data centres with a critical IT load of at least 10 megawatts, and for major cloud providers with an average annual revenue from Singapore users of at least S$100 million. Licensees would be required to implement specific security and resilience requirements such as security risk management, business continuity

[4] For details, see CSA, Cybersecurity Certification Centre at www.csa.gov.sg/our-programmes/certification-and-labelling-schemes/ (Last accessed: 24 September 2026).

[5] For details, see CSA, Cyber Essentials at www.csa.gov.sg/our-programmes/support-for-enterprises/sg-cyber-safe-programme/cybersecurity-certification-for-organisations/cyber-essentials/ (Last accessed: 24 September 2026).

[6] For details, see CSA, Cyber Trust www.csa.gov.sg/our-programmes/support-for-enterprises/sg-cyber-safe-programme/cybersecurity-certification-for-organisations/cyber-trust/#d0b1ad2f10a5cb2d799205d45da5a638 (Last accessed: 24 September 2026).

and disaster recovery measures, and incident reporting (DIB section 15) (Ministry of Digital Development and Information, 2026a).

The Bill comes on the heels of IMDA's 2025 Advisory Guidelines (AGs) for resilience and security of Gloud Services and Data Centres. The AGs were developed in consultation with key CSPs and data centre operators to enhance the resilience and security of their services to minimise service disruptions, which can adversely impact Singapore's economy and society. The AGs set out industry best practices for managing risks ranging from technical architecture misconfigurations and physical hazards to cyber attacks. Key recommended practices include risk assessments, business continuity planning, and cybersecurity measures (IMDA, 2026a).

Partnership Approach

Multi-stakeholder engagement and partnership is integral to Singapore's model. As observed above, the statutory frameworks and codes were developed in consultation with industry stakeholders and users, extending to co-creation in relation to the CSP companion guides. This is consistent with the Singapore Government's long-standing view that cybersecurity is a collective or shared responsibility and multistakeholder engagement and partnership is essential (CSA, 2021a; Teo, 2026a).

Other examples of public-private partnerships outside of the regulatory sphere include the industry-led Cyber Resilience Centre, which is supported by the CSA. The Centre provides health checks and recovery assistance for small enterprises and offers up to co-funding for advisory services under its CISO (Chief Information Security Officer)-as-a-Service programme (Singapore Business Federation, 2026; CSA, 2025b). Additionally, CSA's Operational Technology (OT) Cybersecurity Masterplan 2024, a strategic blueprint to enhance the technical cybersecurity capabilities and competencies of Singapore's OT sector, was co-created with over 60 organisations (including from consulting and professional services organisations, Institutes of Higher Learning & training organisations, CI owners, relevant government agencies, OT Original Equipment Manufacturers OEMs, system integrators, cybersecurity solution providers, cloud service providers, and OT Small and Medium-Sized Enterprises) and launched with commitments from 14 equipment manufacturers and solution providers (CSA, 2024c). These practical-oriented initiatives reflect the logic that Singapore's "collective cyber resilience is only as strong as our weakest link" (MDDI, 2026b), and the necessity of integrating cybersecurity awareness across private-public partnerships to help foster a "proactive defence posture" (Sandi and van den Berg).

4. Thematic Challenges

Supply Chains and Interconnected Systems

Supply chain compromise is treated as a systemic risk rather than a vendor-management issue. This shift can be observed from the amendments to the CA in 2024 that extend reporting duties to supplier-controlled systems, Cyber Trust Mark requirements for auditors and service providers (CSA, 2026b), and CCoP obligations extending to interconnected systems as discussed above.

Singapore also contributes to efforts to enhance supply chain cybersecurity on the international front. Singapore and the United Kingdom, as co-chairs of the Counter Ransomware Initiative, led efforts the development of Guidance for Organisations to Build Supply Chain Resilience Against Ransomware, providing practical recommendations to help organisations strengthen supply chain resilience (UK Home Office, 2025; CSA, 2026l).

Operational Technology (OT)

The emergence of frontier technology is also reshaping the OT[7] threat landscape, allowing attackers without OT expertise to target critical infrastructure in unprecedented ways. A more structured threat ecosystem is also emerging. One group first gains access through information technology systems to steal network diagrams or process configurations, which lays the foundation for a second more specialised OT operations (Teo, 2026b).

CSA's updated OT Cybersecurity Masterplan 2024 seeks to enhance the technical cybersecurity capabilities and competencies of Singapore's OT sector given the evolving nature of cyber-OT operations, such as promoting "Secure-by-Deployment" principles across the entire lifecycle of OT systems (CSA, 2024b). The updated Masterplan contain updates to enhance the OT cybersecurity talent pipeline; improve information sharing and reporting, uplift OT cybersecurity resilience beyond CII, and establish an OT Cybersecurity Centre of Excellence (CSA, 2024c).

Artificial Intelligence, Agentic Systems and Quantum Computing

The technologies driving Singapore's digital economy strategy are also among the most acute sources of risk, which explains why technology governance sits at the centre of the model. On AI governance, commentators have observed that Singapore relies on non-binding frameworks rather than a single dedicated statute. These frameworks are situated within Singapore's wider digital governance regime that prioritises trust and security as well as a commitment to trusted digital infrastructure (Yee & Kim, 2026).

[7] OT refers to systems that control physical processes.

The increasing deployment of AI systems have heightened concerns over the associated cybersecurity risks given their vulnerabilities to adversarial attacks and other cybersecurity risks. Agentic AI introduces new risks with greater potential for impact given their capabilities and access to tools and data. This highlights the importance of AI systems being secure by design and secure by default. In May 2026, in a somewhat unusual move precipitated by recent developments in fronter AI, the CoC and Chief Executive of CSA issued an open letter to boards on the cybersecurity implications of advances in frontier AI. The letter emphasized the need for board-level and CEO attention to the changed cyber risk environment and need to review their cybersecurity risk posture (CSA, 2026c).

To assist organisations to adopt AI in a secure manner, and motivated by the recognition that users need to trust AI functions as originally intended, CSA issued Guidelines and a Companion Guide on Securing AI Systems in late 2024 (CSA, 2024a), followed by an Addendum to the Guidelines and Companion Guide on securing agentic AI systems in June 2026 (CSA, 2026k). The Guidelines identified potential threats and risks in AI systems such as supply chain attacks and Adversarial Machine Learning, set out foundational security principles to secure the use of AI throughout its lifecycle, and highlighted best practices to protect AI systems. The Guidelines further recommend a holistic approach for the identification and mitigation of security risks. The Companion Guide, described as a community driven resource, was developed in collaboration with AI and cybersecurity practitioners. It provides non-prescriptive practical measures and controls that system owners may consider in implementing the Guidelines. Both documents are "living resources" that will be continually updated to account for technological developments (CSA, 2024a).

The Addendum outlines how risks can be identified and assessed on the capabilities of the agentic AI system, and sets out recommended practical measures organisations may consider when securing their adoption of agentic AI systems. The Addendum was similarly developed in collaboration with industry, government and international partners, in line with Singapore's partnership approach as discussed above.

As regards the threat posed by quantum computing to cryptographic systems that secure online transactions, communications and critical services, CSA, IMDA and GovtTech, in consultation with industry partners, jointly published a Quantum-Safe Migration Handbook and Quantum Readiness Index in July 2026. This seeks to addresses the threat proactively, notwithstanding its uncertain timeline. Emphasizing the importance of early preparation for migration, especially for CII owners and government agencies, companies can use these resources to self-assess their current readiness level, identify gaps to be addressed, and use the Handbook for detailed guidance and practical consid-

erations (CSA, 2026j). The Handbook is expressly stated to be non-mandatory, prescriptive or exhaustive (CSA, 2026g).

Foreign Interference and Electoral Integrity

Cybersecurity extends to preserving electoral integrity and countering hostile information campaigns. For the purpose of this chapter, the most relevant legislative framework is the Foreign Interference (Countermeasures) Act 2021 (FICA) (Singapore Statutes Online, 2021) and the Elections (Integrity of Online Advertising) (Amendment) Act 2024 (Singapore Statutes Online, 2024b).

The former targets hostile information campaign (Ministry of Home Affairs, 2026) while the latter targets digitally manipulated content during elections, including AI-generated deepfakes, and empowers the Returning Officer to issue corrective directions (Ministry of Digital Development and Information, 2024).[8] Under the Cybersecurity (Amendment) Act 2024, election systems may additionally be designated as systems of temporary cybersecurity concern (see above).

Assessment

Collectively, the preceding review reveals a digital and AI governance approach that has been described as a pro-innovation, framework-driven governance model, encompassing a network of voluntary, principle-based guidance, sector-specific rules, strong data protection foundations, international interoperability and practical tools and incentives, rather than a single omnibus statute (Yee & Kim, 2026). The accompanying risk-based mechanisms also include market facing certification as well as pre-legislative frameworks tailored to emerging technology. This approach allows for timely and responsive adaptation to technology developments and the associated accelerating and increasingly complex threat landscape. The quote below from CSA's former chief executive best encapsulates this philosophy:

> We need to lock down, find first, and fix fast. We must tighten up and harden systems before threat actors can find footholds. Discover vulnerabilities before they are exploited. And when gaps are found, close them faster than threat actors can act. This cycle should be continuous, rather than point-in-time checks. We are working closely with global partners and industry to ensure that fast-evolving technologies strengthen, rather than undermine, our collective security. Together – industry, government, and citizens – we can build a future where digital innovation thrives in tandem with trust and security (CSA, 2026f).

[8] The press release announced a new Code of Practice that would require social media services to implement measures to prevent and counter the abuse of digitally manipulated online content on their service. This was not available at the time of writing.

Digital Defence

A holistic whole-of-nation approach can be discerned from the addition of Digital Defence as sixth pillar of Total Defence in 2019, where it operates alongside military, civil, economic, social and psychological defence to guard against threats from the digital domain (Ministry of Communications and Information & Ministry of Defence, 2019). Recognising that cyber resilience is partly behavioural, digital defence asks individuals to practise good cybersecurity habits, guard against fake news and disinformation, and consider the impact of their online actions on the community (Ministry of Defence, 2019).[9]

As part of efforts to strengthen whole of society resilience, pilot National Simulated Scams Exercise were launched in March 2026. The exercise simulated AI-enabled government official impersonation scam calls and served to increase public awareness of and response to emerging scam tactics (CSA, 2026f: 11). Additionally, a guidance note on what to do during a digital connectivity disruption was published, furthering the Total Defence approach (CSA, 2026i).[10]

5. Regional Architecture and Norms Entrepreneurship

The third strategic pillar reflects a recognition that Singapore's security cannot be secured within its borders alone. This perspective should also be viewed against ASEAN's shared interest in securing the ICT space for the digital economy as gleaned from ASEAN's Digital Masterplan 2025 Desired Outcome 3 – "the delivery of trusted digital services and the prevention of consumer harm" (ASEAN, 2021) and the ASEAN Digital Masterplan 2030 Desired Outcome 3 – "Empower, Safe and Inclusion ASEAN Digital Society" (ASEAN, 2026).

At the 32nd ASEAN Summit in 2018, under Singapore's chairmanship, ASEAN leaders tasked relevant ASEAN Ministers to among others, identify a concrete list of voluntary, practical norms of State behaviour in cyberspace that ASEAN can work towards adopting and implementing, taking reference from the voluntary norms recommended in the 2015 Report of the United Nations Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security (UNGGE) (ASEAN, 2018). Subsequently, at the third ASEAN Ministerial Conference on Cybersecurity (AMCC) in September 2018, ASEAN became the first regional

[9] See also Total Defence, "What is Digital Defence?" infographic, available online at isomer-user-content.by.gov.sg/129/bb2158ef-9ed6-4ff0-9174-652461d33e36/what-is-digital-defence.pdf (Last accessed: 24 September 2026).

[10] Full text of guidance can be found at www.imda.gov.sg/user-and-set-up-guides/mobile-and-broadband/what-to-do-during-a-digital-connectivity-disruption (Last accessed 24 September 2026).

grouping to subscribe in principle to the UNGGE's eleven voluntary norms of responsible state behaviour, reaffirming "the importance of a rules-based cyberspace as an enabler of economic progress and better of living standards" (CSA, 2018b). In addition to this, a working group was established to chart a long-term roadmap for implementing the norms (ASEAN, 2022).

Singapore's Role in Implementing the UN Norms: the Norms Implementation Checklist

As a starting point, Singapore's 2021 national position on international law affirms that international law, in particular the UN Charter, applies in cyberspace and stresses sovereignty and non-intervention as principles of fundamental importance to small states (United Nations, 2021).

Turing to the UNGGE norms, translating the eleven voluntary norms into a practical tool has been a distinctive Singapore contribution. At the fifth Singapore International Cyber Week in 2020, Singapore and the UN jointly announced their cooperation to develop a Norms Implementation Checklist (NIC) that would outline a list of steps that countries need to take to implement the 11 UNGGE norms (CSA, 2020).

Singapore and Malaysia ultimately co-led the completion of the NIC with the support of the UN Office for Disarmament Affairs (UNODA). This is a significant achievement, as it is the first regional checklist of its kind and serves as a reference both for ASEAN states and beyond in support of global efforts towards a safer and more secure cyberspace (CSA, 2024d). The NIC was eventually endorsed at the 5th ASEAN Cybersecurity Coordinating Committee Meeting chaired by Singapore, and announced by Singapore's Minister Josephine Teo at the 5th ASEAN Digital Ministers' Meeting (ADGMIN) on 16 January 2025 (Teo, 2025).

The NIC (ASEAN, 2025) comprises a set of actions that states can consider and follow in implementing each of the eleven UN norms of responsible state behaviour, with the actionable items for every norm organised across five pillars: policy, operation, technical, legal, and diplomacy. Additionally, it is supported with suggested capacity-building activities (Industrial Cyber, 2024).

The conclusion and endorsement of the NIC was significant as Singapore was elected as Chair of the 2021-2025 UN Open Ended Working Group on ICT in the Context of International Security (OEWG) (Permanent Mission of Singapore to the United Nations, 2021). The presentation of ASEAN's norms implementation framework to the OEWG as a regional initiative to support the OEWG's norm-based framework allowed ASEAN to share their ideas on how the UNGGE norms can be implemented domestically as a practical matter. As intimated in the introduction to the NIC, it aims "to provide actionable steps for all States to consider, including small States with limited capacities" (ASEAN, 2025). The

NIC focuses on pragmatic implementation rather than ideological approaches to global governance. States have the flexibility to calibrate the steps according to "respective national priorities and capacities" with a view to converging on a common understanding of expected State conduct in cyberspace, "thereby supporting international peace and security in cyberspace" (ASEAN, 2025).

Alongside the NIC was a parallel effort funded by the United Kingdom and Australian governments and delivered by the Australian Strategic Policy Institute, which led to the publication of "The UN norms of responsible state behaviour in cyberspace Guidance on implementation for Member States of ASEAN" (Australian Strategic Policy Institute, 2022). This document contains additional guidance drawn from the 2021 UNGGE report, illustrative questions for policymakers, and examples of observed good practice from ASEAN states. The coexistence of these two efforts, one Singapore-led and government-to-government, one externally funded and delivered by a think tank, illustrates a wider pattern in ASEAN norms implementation: the region relies on a mix of member-state initiative and international partner support to convert a voluntary, non-binding UN framework into operational guidance that individual states can act on. As Chair of the OEWG, Ambassador Burhan Gafoor, Permanent Representative of Singapore to the UN, successfully steered the OEWG to the adoption of its Final Report by consensus in July 2025 where the work on the voluntary norms continues in the successor Global Mechanism process.

Admittedly, limits exist regarding consensus-based regionalism. Tran concludes that Southeast Asia has made considerable progress across the four pillars of the UN framework but must ensure that diplomatic engagement translates into operational reality (Tran, 2025). Voluntary norms still depend on political will and on the capacity of the least resourced members, and the approach that confers legitimacy in ASEAN, through consultation and flexibility, does not by itself produce enforceability.

Apart from the NIC, Singapore continues to contribute to and support ASEAN's cybersecurity coordination, information sharing and capacity-building. Singapore led the establishment of, as well as funding and hosting the physical facilities of the ASEAN Regional Computer Emergency Response Team (CSA, 2022; CSA 2024d). In addition, Singapore launched the ASEAN-Singapore Cybersecurity Centre of Excellence to strengthen regional cybersecurity capacity (CSA, 2021c) and also hosts the ADMM Cybersecurity and Information Centre of Excellence (Fitraini, 2023)[11].

[11] The ACICE was established to focus on confidence-building measures, enhancing information-sharing, and capacity building among ASEAN defence establishments. The ACICE aims to enhance multilateral cooperation among ASEAN defence establishments against cyberattacks, disinformation and misinformation.

Recalling ASEAN's shared interest in securing the ICT space for the digital economy, Singapore also played an active role in the negotiations on the ASEAN Digital Economy Framework Agreement (DEFA). The DEFA is expected to be signed in November 2026. This will mark the world's first comprehensive, region-wide agreement dedicated to the digital economy (Ministry of Trade and Industry, 2026c); designed to harmonize digital trade rules, enable trusted cross-border data flows, and establish coherent regulations for paperless trading, cybersecurity and digital payments and electronic invoicing across ASEAN (Feingold and Pfister, 2025).

6. Conclusion

Singapore's model combines features that are individually familiar and unique in combination.

Singapore exercises sovereignty by choosing where control is essential, favouring trusted cross-border data flows and resisting unjustified data localisation. Regional and global norm-building are also leveraged on as an extension of national resilience.

Its domestic governance framework is anchored in statute, with operative detail sitting in codes and certification schemes, advisories and guidelines that can be revised within months rather than years. It is risk-based and calibrated, scaling obligations to systemic importance and balancing considerations of security and technology risks, innovation and economic opportunities, as well as business compliance cost. The model provides regulatory certainty while also demonstrating flexibility and nimbleness where needed. Finally, broad and deep multi-stakeholder is evident, with non-government stakeholders treated as partners and in some cases, co-creators.

For other small states, Singapore's model provides a reference point rather than a template. Its conditions include high state capacity, high levels of trust between government and industry, and a scale that allows close engagement with regulated entities. Several lessons are nonetheless salient: (a) anchor the regime in law but keep operational details flexible; (b) differentiate obligations by systemic importance; (c) use certification to extend baselines through supply chains; (d) invest in exercises and partnerships; and (e) leverage regional leadership to build upon and operationalise the norms that protect an open economy.

The upcoming strategy review, code for cloud services, and Singapore's ASEAN chairmanship in 2027 is expected to provide more indications as to Singapore's future strategic priorities and initiatives in the face of rapid technological advancements, and an increasingly complex and evolving cyber threat landscape.

References

ASEAN. 2018. *ASEAN Leaders' Statement on Cybersecurity Cooperation.* Available online at asean.org/asean-leaders-statement-on-cybersecurity-cooperation/ (Last accessed: 30 September 2026).

ASEAN. 2021. *Digital Masterplan 2025.* Available online at https://asean.org/wp-content/uploads/2021/09/ASEAN-Digital-Masterplan-EDITED.pdf (Last accessed: 30 September 2026).

ASEAN. 2022. *ASEAN Cybersecurity Cooperation Strategy (2021–2025).* Available online at asean.org/wp-content/uploads/2022/02/01-ASEAN-Cybersecurity-Cooperation-Paper-2021-2025_final-23-0122.pdf (Last accessed: 21 September 2026).

ASEAN. 2025. *ASEAN Checklist for the Implementation of the Norms of Responsible State Behaviour in Cyberspace*. Available online at asean.org/wp-content/uploads/2025/02/ASEAN_checklist_print.pdf (Last accessed: 30 September 2026).

ASEAN. 2026. *ASEAN Digital Masterplan 2030.* Available at https://asean.org/wp-content/uploads/2026/01/ADM-2030.pdf (Last accessed: 30 September 2026).

Australian Strategic Policy Institute. 2022. *The UN norms of responsible state behaviour in cyberspace – Guidance on implementation for Member States of ASEAN.* Available online at documents.unoda.org/wp-content/uploads/2022/03/The-UN-norms-of-responsible-state-behaviour-in-cyberspace.pdf (Last accessed: 30 September 2026).

Cyber Security Agency of Singapore. 2016. *ASEAN Member States Call for Tighter Cybersecurity Coordination in ASEAN (*Press Release). Available online at www.csa.gov.sg/news-events/press-releases/asean-member-states-call-for-tighter-cybersecurity-coordination-in-asean/ (Last accessed: 21 September 2026).

Cyber Security Agency of Singapore. 2018a. *Speech by DPM Teo Chee Hean At the Opening of the Third Singapore International Cyber Week.* Available online at www.csa.gov.sg/news-events/speeches/singapore-international-cyber-week-2018-opening-ceremony/ (Last accessed : 24 September 2026).

Cyber Security Agency of Singapore. 2018b. *ASEAN Member States Agree to Strengthen Cyber Coordination and Capacity-Building Efforts* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/amcc-2018/ (Last accessed: 30 September 2026).

Cyber Security Agency of Singapore. 2020. *Singapore International Cyber Week 2020 – Highlights and Testimonials* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/sicw-2020-highlights-and-testimonials/ (Last accessed: 30 September 2026).

Cyber Security Agency of Singapore. 2021a. *Singapore Updates National Cybersecurity Strategy* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/singapore-updates-national-cybersecurity-strategy/ (Last accessed: 24 September 2026).

Cyber Security Agency of Singapore. 2021b. *The Singapore Cybersecurity Strategy 2021.* Available online at www.csa.gov.sg/resources/publications/the-singapore-cybersecurity-strategy-2021/ (Last accessed: 22 September 2026).

Cyber Security Agency of Singapore. 2021c. *ASEAN-Singapore Cybersecurity Centre of Excellence* (Press Release). Available online at www.csa.gov.sg/news-events/

press-releases/asean-singapore-cybersecurity-centre-of-excellence/ (Last accessed: 30 September 2026).

Cyber Security Agency of Singapore. 2022. *Establishment of ASEAN Regional Computer Emergency Response Team (CERT)* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/establishment-of-asean-regional-computer-emergency-response-team/ (Last accessed: 30 September 2026).

Cyber Security Agency of Singapore. 2024a. *Guidelines and Companion Guide on Securing AI Systems.* Available online at www.csa.gov.sg/resources/publications/guidelines-and-companion-guide-on-securing-ai-systems/ (Last accessed: 24 September 2026).

Cyber Security Agency of Singapore. 2024b. *Singapore's Operational Technology Cybersecurity Masterplan 2024*. Available online at www.csa.gov.sg/resources/publications/singapore-s-operational-technology-cybersecurity-masterplan-2024/ (Last accessed: 24 September 2026).

Cyber Security Agency of Singapore. 2024c. *Singapore Updates Operational Technology Cybersecurity Masterplan* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/singapore-updates-operational-technology-cybersecurity-masterplan/ (Last accessed: 21 September 2026).

Cyber Security Agency of Singapore. 2024d. *Singapore and ASEAN Member States Deepen Commitment to Enhance Collective Cybersecurity in the Region* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/singapore-and-asean-member-states-deepen-commitment-to-enhance-collective-cybersecurity-in-the-region/ (Last accessed: 30 September 2026).

Cyber Security Agency of Singapore. 2025a. *Opening Speech by Mr K Shanmugam, Coordinating Minister for National Security and Minister for Home Affairs, at CSA's 10th Anniversary Dinner.* Available online at www.csa.gov.sg/news-events/speeches/opening-speech-by-mr-k-shanmugam--coordinating-minister-for-national-security-and-minister-for-home-affairs--at-csa-10th-anniversary-dinner/ (Last accessed: 21 September 2026).

Cyber Security Agency of Singapore. 2025b. *Plans for Establishment of Cyber Resilience Centre to Provide Support To Organisations* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/plans-for-establishment-of-cyber-resilience-centre-to-provide-support-to-organisations/ (Last accessed: 24 September 2026).

Cyber Security Agency of Singapore. 2025c. *Provisions in the Cybersecurity (Amendment) Act to Come Into Force on 31 October 2025* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/provisions-in-the-cybersecurity--amendment--act-to-come-into-force-on-31-october-2025/ (Last accessed: 23 September 2026).

Cyber Security Agency of Singapore. 2026a. *Advisory on Risks Associated with Frontier AI Models*. Available online at www.csa.gov.sg/alerts-and-advisories/advisories/ad-2026-004/ (Last accessed: 22 September 2026).

Cyber Security Agency of Singapore. 2026b. *CSA to Raise Cybersecurity Standards for Critical Information Infrastructure Owners* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/csa-to-raise-cybersecurity-standards-for-critical-information-infrastructure-owners/ (Last accessed: 24 September 2026).

Cyber Security Agency of Singapore. 2026c. *Commissioner of Cybersecurity and Chief Executive of Cyber Security Agency of Singapore's open letter to boards on cyber security implications of frontier AI.* Available online at www.sid.org.sg/Web/Web/About/News-and-Press/Regulatory-Announcements/Commissioner_of_Cybersecurity_and_Chief_Executive.aspx (Last accessed: 24 September 2026).

Cyber Security Agency of Singapore. 2026d. *Cybersecurity Code of Practice for Critical Information Infrastructure 2026.* Available online at www.csa.gov.sg/legislation/codes-of-practice/ (Last accessed: 24 September 2026).

Cyber Security Agency of Singapore. 2026e. *Cybersecurity Code of Practice for Critical Information Infrastructure to be Updated to Address APT and AI-enabled Threats* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/cybersecurity-code-of-practice-for-critical-information-infrastructure-to-be-updated-to-address-apt-and-ai-enabled-threats/ (Last accessed: 24 September 2026).

Cyber Security Agency of Singapore. 2026f. *CSA's Initiatives to Strengthen Singapore's Cyber Defences Amid an AI-Driven Threat Landscape* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/csa-s-initiatives-to-strengthen-singapore-s-cyber-defences-amid-an-ai-driven-threat-landscape/ (Last accessed: 22 September 2026).

Cyber Security Agency of Singapore. 2026g. *Singapore Cyber Threat Landscape 2025/2026*. Available online at isomer-user-content.by.gov.sg/36/627d81a0-3e85-446c-bddb-587ba9df42f4/Singapore%20Cyber%20Landscape%202025_2026.pdf (Last accessed: 24 September 2026).

Cyber Security Agency of Singapore. 2026h. *Largest Multi-Agency Cyber Operation Mounted to Counter Threat Posed by Advanced Persistent Threat (APT) Actor UNC3886 to Singapore's Telecommunications Sector* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/largest-multi-agency-cyber-operation-mounted-to-counter-threat-posed-by-advanced-persistent-threat--apt--actor-unc3886-to-singapore-s-telecommunications-sector/ (Last accessed: 21 September 2026).

Cyber Security Agency of Singapore. 2026i. *Opening Speech by Minister Josephine Teo at the Launch of Exercise SG Ready 2026 and Total Defence Commemoration Event*. Available online at www.csa.gov.sg/news-events/speeches/opening-speech-by-minister-josephine-teo-at-the-launch-of-exercise-sg-ready-2026-and-total-defence-commemoration-event-on-1-feb-2026/ (Last accessed: 24 September 2026).

Cyber Security Agency of Singapore. 2026j. *Quantum-Safe Migration Handbook and Quantum Readiness Index*. Available online at www.csa.gov.sg/resources/publications/quantum-safe-handbook-and-quantum-readiness-index/ (Last accessed 24 September 2026).

Cyber Security Agency of Singapore. 2026k. *Securing Agentic AI – An Addendum to the Guidelines and Companion Guide on Securing AI Systems*. Available online at www.csa.gov.sg/resources/publications/addendum-on-securing-ai-systems/ (Last accessed 24 September 2026).

Cyber Security Agency of Singapore. 2026l. *Singapore Cyber Landscape 2025/2026.* Available online at https://www.csa.gov.sg/resources/publications/singapore-cyber-landscape-2025-2026/ (Last accessed: 22 September 2026).

Economic Strategy Review Committees. 2026. *Economic Strategy Review (ESR) Final Report.* Available online at https://www.gov.sg/features/esr/ (Last accessed: 1 October 2026).

Feingold, S. and Pfister, A. 2025. *ASEAN Takes Major Step toward Landmark Digital Economy Pact*. Available online at https://www.weforum.org/stories/economic-growth/asean-defa-digital-economy-pact-negotiations/ (Last accessed: 30 September 2026).

Fitriani. 2023. *New structures for ASEAN cyber defence and information sharing*. Available online at www.iiss.org/online-analysis/online-analysis/2023/09/new-structures-for-asean-cyber-defence-and-information-sharing/ (Last accessed: 30 September 2026).

Global Financial Centres Index (GFCI), Z/Yen Group and China Development Institute. 2026. *The Global Financial Centres Index 40 (GFCI 40)*. Available online at longfinance.net/programmes/financial-centre-futures/global-financial-centres-index/ (Last accessed: 21 September 2026).

Industrial Cyber. 2024. *ASEAN Ministers Unite to Bolster Cyber Resilience at 9th Cybersecurity Conference.* Available online at industrialcyber.co/news/asean-ministers-unite-to-bolster-cyber-resilience-at-9th-cybersecurity-conference/ (Last accessed: 21 September 2026).

Infocomm Media Development Authority. 2025. *Singapore's Digital Economy at 18.6% of GDP, up from 14.9% in 2019; Growth fuelled by accelerating digitalisation and AI adoption across sectors and firms* (Press Release). Available online at www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2025/singapore-digital-economy (Last accessed: 21 September 2026).

Infocomm Media Development Authority. 2026a. *IMDA introduced Advisory Guidelines to enhance the resilience and security of Cloud Services and Data Centres* (Press Release). Available online at www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/advisory-guidelines-of-cloud-services-and-data-centres (Last accessed 24 September 2026).

Infocomm Media Development Authority. 2026b. *Infocomm Media Cyber Security.* Available online at www.imda.gov.sg/regulations-and-licensing-listing/infocomm-media-cyber-security (Last accessed: 23 September 2026).

Menon, R. 2018. *Singapore FinTech – Innovation, Inclusion, Inspiration*. Available online at www.bis.org/review/r181112a.htm (Last accessed: 22 September 2026).

Ministry of Communications and Information, Ministry of Defence. 2019. *Total Defence Strengthened with Addition of Digital Defence as the Sixth Pillar* (Joint News Release). Available online at www.mindef.gov.sg/news-and-events/latest-releases/15feb19_nr (Last accessed: 24 September 2026).

Ministry of Defence. 2019. *Fact Sheet: Digital Defence.* Available online at www.mindef.gov.sg/news-and-events/latest-releases/15feb19_fs/ (Last accessed: 21 September 2026).

Ministry of Digital Development and Information. 2024. *New Legal Measures to Uphold Integrity of Online Advertising during Elections* (Press Release). Available online at www.mddi.gov.sg/newsroom/new-legal-measures-uphold-integrity-online-advertising-elections/ (Last accessed: 21 September 2026).

Ministry of Digital Development and Information. 2026a. *New Digital Infrastructure Bill to Strengthen the Foundations for Singapore's Digital Economy* (Press Release). Available online at www.mddi.gov.sg/newsroom/new-digital-infrastructure-bill-to-strengthen-the-foundations-for-singapore-s-digital-economy/ (Last accessed: 21 September 2026).

Ministry of Digital Development and Information. 2026b. *Transcript Of Keynote Address By Minister Josephine Teo At The Operational Technology Cybersecurity Expert Panel Forum 2026.* Available online at www.mddi.gov.sg/newsroom/transcript-of-keynote-address-by-minister-josephine-teo-at-the-operational-technology-cybersecurity-expert-panel-forum-2026/ (Last accessed: 24 September 2026).

Ministry of Home Affairs. 2026. *Introduction to Foreign Interference (Countermeasures) Act(FICA).* Available online at www.mha.gov.sg/what-we-do/foreign-interference/fica/ (Last accessed: 21 September 2026).

Ministry of Trade and Industry. 2017. *Report of the Committee on the Future Economy: Pioneers of the Next Generation.* Available online at www.meti.gov.sg/newsroom/report-of-the-committee-on-the-future-economy/ (Last accessed: 2 October 2026).

Ministry of Trade and Industry. 2025. *Singapore and the European Free Trade Association Strengthen Economic and Digital Connectivity with the Signing of the European Free Trade Association Digital Economy Agreement* (Press Release). Available online at www.meti.gov.sg/newsroom/singapore-and-the-european-free-trade-association-strengthen-economic-and-digital-connectivity/ (Last accessed: 2 October 2026).

Ministry of Trade and Industry. 2026a. *Digital Economy Agreements.* Available online at www.meti.gov.sg/trade-international-economic-relations/agreements/digital-economy-agreements-dea/#adb912736be5ff7e0a68b9d3a237ab95 (Last accessed: 2 October 2026).

Ministry of Trade and Industry. 2026b. *Speech by Deputy Prime Minister and Minister for Trade and Industry Gan Kim Yong at MTI Committee of Supply Debate 2026*. Available online at www.meti.gov.sg/newsroom/speech-by-deputy-prime-minister-and-minister-for-trade-and-industry-gan-kim-yong-at-mti-committee-of-supply-debate-2026/ (Last accessed: 2 October 2026).

Ministry of Trade and Industry. 2026c. *Singapore and ASEAN Member States Conclude Negotiations on Landmark ASEAN Digital Economy Framework Agreement (DEFA)* (Press Release). Available online www.meti.gov.sg/newsroom/singapore-and-asean-member-states-conclude-negotiations-on-landmark-asean-defa (Last accessed: 2 October 2026).

Monetary Authority of Singapore. 2021. *Guidelines on Risk Management Practices – Technology Risk*. Available at www.mas.gov.sg/regulation/guidelines/technology-risk-management-guidelines (Last accessed: 23 September 2026).

Monetary Authority of Singapore. 2024. *Notice FSM-N05 Technology Risk Management.* Available online at www.mas.gov.sg/regulation/notices/notice-fsm-n05 (Last accessed: 23 September 2026).

Monetary Authority of Singapore. 2026. *Consultation Paper on Proposed Amendments to Notices of Technology Risk Management*. Available online at www.mas.gov.sg/publications/consultations/2026/consultation-paper-on-proposed-amendments-to-notices-on-technology-risk-management (Last accessed: 23 September 2026).

Permanent Mission of Singapore to the United Nations. 2021. *Statement of 29 June 2021.* Available online at newyork-un.mfa.gov.sg/mission-updates/security-council/press-20210629-29-jun-2021/ (Last accessed: 21 September 2026).

Personal Data Protection Commission. n.d. *Regulatory Guidance.* Available online at www.pdpc.gov.sg/organisations/regulations-decisions/regulatory-guidance?page=1&sort=latest (Last accessed: 23 September 2026).

Rajah & Tann Asia. 2026. *Updates on Cybersecurity Code of Practice for Critical Information Infrastructure and Cloud Services*. Available online at www.rajahtannasia.com/viewpoints/updates-on-cybersecurity-code-of-practice-for-critical-information-infrastructure-and-cloud-services/ (Last accessed: 24 September 2026).

Sandi, S., van den Berg, C. 2025. *Cybersecurity mindset and upskilling: Resilience via lifelong learning and security education.* Available online at https://sajim.co.za/index.php/sajim/article/view/2044/3336 (Last accessed: 8 October 2026).

Singapore Business Federation. 2026. *SBF established cyber resilience centre to strengthen cybersecurity support for business* (Press Release). Available online at www.csa.gov.sg/news-events/press-releases/csa-to-invest-50-million-to-uplift-talent-innovation-and-growth-in-singapore-s-cybersecurity-support-for-businesses (Last accessed: 24 September 2026).

Singapore International Cyber Week. 2026. *About Singapore International Cyber Week.* Available online at www.sicw.gov.sg/about-singapore-international-cyber-week/ (Last accessed: 21 September 2026).

Singapore Parliament. 2024. *Second Reading Bills: Cybersecurity (Amendment) Bill* (Official Reports). Available online at https://sprs.parl.gov.sg/search/#/sprs3topic?reportid=bill-692 (Last accessed: 23 September 2026).

Singapore Statutes Online. 2021. *Foreign Interference (Countermeasures) Act 2021*. Available online at sso.agc.gov.sg/Act/FICA2021 (Last accessed 24 September 2026).

Singapore Statutes Online. 2024. *Cybersecurity (Amendment) Act 2024*. Available online at sso.agc.gov.sg/Acts-Supp/19-2024/Published/20240704?DocDate=20240704 (Last accessed: 23 September 2026).

Singapore Statutes Online. 2024b. *Elections (Integrity of Online Advertising) (Amendment) Act 2024.* Available online at sso.agc.gov.sg/Acts-Supp/34-2024/Published/20241121?DocDate=20241121 (Last accessed 24 September 2026).

Singapore Statutes Online. 2025. *Cybersecurity Act 2018 (as amended)*. Available online at sso.agc.gov.sg/Act/CA2018?ValidDate=20251031&Timeline=On (Last accessed: 23 September 2026).

Singapore Statutes Online. 2026a. *Digital Infrastructure Bill (Bill No. 20/2026)*. Available online at sso.agc.gov.sg/Bills-Supp/20-2026/Published/20260908?DocDate=20260908 (Last accessed: 23 September 2026).

Singapore Statutes Online. 2026b. *Health Information Act 2026*. Available at sso.agc.gov.sg/Act/HIA2026/Uncommenced/20260924070832?DocDate=20260212 (Last accessed: 24 September 2026).

Singapore Statutes Online. 2026c. *Personal Data Protection Act 2012 (PDPA)*. Available online at sso.agc.gov.sg/Act/PDPA2012?ProvIds=P16-#pr24- (Last accessed: 23 September 2026).

Tan, K. 2026. *Senior Minister of State, Tan Kiat How, Committee of Supply 2026 Speech, Safeguarding our Digital Space in a Digital Age.* Available online at www.csa.gov.sg/news-events/speeches/senior-minister-of-state-tan-kiat-how-committee-of-supply-2026-speech/ (Last accessed: 24 September 2026).

Teo, J. 2022. *Speech by Minister Josephine Teo at the Global Technology Summit.* Available online at https://www.mddi.gov.sg/newsroom/minister-josephine-teo-at-the-gobal-technology-summit/ (Last accessed: 22 September 2026).

Teo, J. 2025. *Opening Remarks at the 5th ASEAN Digital Ministers' Meeting*. Available online at www.mddi.gov.sg/newsroom/opening-remarks-by-minister-josephine-teo-at-the-5th-adgmin/ (Last accessed: 21 September 2026).

Teo, J. 2026a. *Opening Speech at the Operation CYBER GUARDIAN Engagement Event for Cyber Defenders*. Available online at www.mddi.gov.sg/newsroom/opening-speech-by-minister-josephine-teo-at-the-operation-cyber-guardian-engagement-event-for-cyber-defenders/ (Last accessed: 21 September 2026).

Teo, J. 2026b. *Keynote Address at the Operational Technology Cybersecurity Expert Panel Forum 2026*. Available online at www.mddi.gov.sg/newsroom/transcript-of-keynote-address-by-minister-josephine-teo-at-the-operational-technology-cybersecurity-expert-panel-forum-2026/ (Last accessed: 24 September 2026).

Tran, B. 2025. *Southeast Asia's Cybersecurity Trajectory beyond the UN OEWG.* Available online at rsis.edu.sg/rsis-publication/idss/ip25108-southeast-asias-cybersecurity-trajectory-beyond-the-un-oewg/ (Last accessed: 21 September 2026).

UK Home Office. 2025. *Guidance for organisations to build supply chain resilience against ransomware* (Guidance). Available online at www.gov.uk/government/publications/supply-chain-resilience-against-ransomware/guidance-for-organisations-to-build-supply-chain-resilience-against-ransomware (Last accessed: 24 September 2026).

United Nations. 2021. *Official Compendium of Voluntary National Contributions on the Subject of How International Law Applies to the Use of Information and Communications Technologies by States, UN Doc.* Available online at digitallibrary.un.org/record/3933543/files/A_76_136-EN.pdf (Last accessed: 21 September 2026).

Voice of Indonesia. 2026. *ASEAN Targets DEFA Signing at 2026 Summit.* Available online at rri.co.id/en/business/2398715/asean-targets-defa-signing-at-2026-summit (Last accessed: 21 September 2026).

Wong Partnership. 2021. *MAS Releases Revised Technology Risk Management Guidelines.* Available online at https://www.wongpartnership.com/upload/medias/KnowledgeInsight/document/15308/LegisWatch_MASReleasesRevisedTechnologyRiskManagementGuidelines.PDF (Last accessed: 23 September 2026).

Yee, L., Kim, S. 2026. *Singapore's Digital & AI Governance: A Pro-Innovation, Framework-Driven Model*. Available online at blogs.duanemorris.com/duanemorrisandselvam/2026/03/03/singapores-digital-ai-governance-a-pro-innovation-framework-driven-model/ (Last accessed: 24 September 2026).

MARITIME CYBERSECURITY AND RESILIENCE: FROM NIS2 COMPLIANCE TO THE PROTECTION OF CRITICAL MARITIME AND UNDERWATER INFRASTRUCTURE

Stefano Melina
Italian Coast Guard, Italy

Abstract
The maritime transport sector serves as an indispensable pillar of global commerce, carrying over 80% of international trade. However, escalating geopolitical competition, persistent chokepoint vulnerabilities, and rapid digital transformation have introduced systemic vulnerabilities across maritime routes and supply chains. Modern shipping and port operations increasingly rely on interconnected digital ecosystems – such as the European Maritime Single Window environment (EMSWe), Vessel Traffic Services (VTS), and industrial control systems – which significantly expand the cyberattack surface. Furthermore, the emergence of Maritime Autonomous Surface Ships (MASS) and the operational deployment of unmanned aerial, surface, and underwater systems (RPAS, USVs, AUVs, and ROVs) generate complex cyber-physical risks, including GNSS spoofing, compromise of remote communication links, and cognitive overload among human operators in command-and-control centres.
This paper provides a structured analysis of maritime cybersecurity and operational resilience, examining the interrelated requirements of Directive (EU) 2022/2555 (NIS2) and Directive (EU) 2022/2557 (CER), alongside sectoral benchmarks such as the IMO MASS Code (MSC.595(111)) and Circular No. 177/2025 issued by the Italian Coast Guard and the Ministry of Infrastructure and Transport. Drawing on operational testbeds – notably the 5G MASS project in the Port of Livorno – and examining the protection of critical subsea cables and energy pipelines, this study demonstrates that cybersecurity can no longer be treated as an isolated information technology domain. Instead, ensuring operational continuity, navigational safety, and the protection of essential underwater infrastructure requires an integrated all-hazards governance framework. This approach harmonises cyber risk management, physical asset protection, workforce competence, and institutional cooperation between public- and private-sector stakeholders across the entire maritime ecosystem.

1. Introduction

The Strategic Role of Maritime Transport

In the current international geopolitical environment, characterised by armed conflicts, regional tensions and growing competition for control over major trade routes, the transport sector is assuming increasing strategic importance. Within this context, maritime transport represents one of the pillars of the global econo-

my, serving as the principal means through which international trade in goods is conducted. According to the Review of Maritime Transport 2025, published by the United Nations Conference on Trade and Development (UNCTAD), more than 80% of the volume of international trade in goods is carried by sea. The same report highlights that, following growth of 2.2% in 2024, global maritime trade is expected to slow to 0.5% in 2025, partly as a result of geopolitical tensions and the reconfiguration of shipping routes (UNCTAD, 2025).

The Vulnerability of Maritime Routes and Its Impact on the Global Economy

The strategic importance of the maritime sector becomes particularly evident during crises affecting major maritime transit points, commonly referred to as maritime chokepoints. Events involving the Strait of Hormuz and the Red Sea corridor, through the Bab el-Mandeb Strait, demonstrate how the security and continuity of navigation can directly affect international economic stability. The Strait of Hormuz, in particular, is an essential passage for global energy trade: according to UNCTAD analyses published in March 2026, approximately one quarter of global seaborne oil trade passes through this corridor (UNCTAD, 2026). Disruptions to navigation, attacks against vessels and the need to adopt alternative routes lead to longer transit times and higher fuel and insurance costs, producing cascading effects across supply chains. These consequences extend beyond maritime operators to the entire production system, affecting energy availability, food supplies, commodity prices and, ultimately, the cost of living for citizens.

Digitalization of the Maritime Sector and new Technological Dependencies

At the same time, the maritime sector is undergoing a profound digital transformation, driven by the introduction of integrated platforms, automated traffic management systems and electronic information exchange mechanisms. The objective is to improve operational efficiency, reduce vessel turnaround times in ports, optimise cargo handling and ensure greater interoperability among maritime and port stakeholders. Digital systems for port call planning, port operations management and data sharing enable the coordination of activities involving shipping companies, maritime authorities, terminal operators, logistics providers and essential service providers. This evolution, however, creates increasing dependence on information technology infrastructure and communication networks. The availability, integrity and reliability of digital systems therefore become essential conditions for ensuring the continuity of maritime and port operations.

EMSWe and the Digitalisation of Port Procedures

A significant example of this process is the European Maritime Single Window environment (EMSWe), established by Regulation (EU) 2019/1239 and

applicable from 15 August 2025 (EU, 2019). The system aims to harmonise ship reporting obligations for vessels calling at ports within the European Union through a common data set, standardised interfaces and interconnected national maritime single windows. EMSWe helps simplify administrative formalities, facilitate structured information exchange and integrate maritime transport into the European multimodal logistics chain. These tools are complemented by port community management platforms, access control systems, surveillance solutions and Vessel Traffic Services (VTS). Integration among these systems increases overall efficiency but also expands the cyberattack surface, making it necessary to adopt a security approach that considers not only individual devices but the entire interconnected digital ecosystem.

Autonomous Ships, Remote Control Centres, and new Operational Scenarios

Technological transformation directly affects vessels as well. The development of Maritime Autonomous Surface Ships (MASS), namely autonomous or remotely operated surface ships, introduces new operational models in which navigation systems, sensors, command platforms and communications play a decisive role. In this context, Remote Operation Centres (ROCs) may perform supervisory, control and decision-support functions, including in situations where no crew, or only a reduced crew, is present on board. On 22 May 2026, the International Maritime Organization (IMO) adopted the first International Code of Safety for Maritime Autonomous Surface Ships (MASS Code), initially applicable on a voluntary basis from 1 July 2026, with a subsequent pathway towards a mandatory code. The transition towards increasingly autonomous and connected navigation creates new protection requirements: the compromise of control systems, ship-to-shore links or remote management platforms could affect not only cybersecurity but also navigational safety, the protection of human life at sea and the marine environment.

The NIS2 Directive and Cybersecurity Resilience in the Maritime Sector

Within this scenario, cybersecurity assumes strategic importance and can no longer be regarded solely as a matter of protecting information systems. It constitutes an essential component of the operational resilience of the infrastructure and services upon which maritime transport depends. At European level, the process of strengthening the security of network and information systems began with Directive (EU) 2016/1148, commonly known as NIS1, and continued with Directive (EU) 2022/2555, known as NIS2. The latter expands and strengthens the framework of obligations concerning risk management, supply chain security, incident handling and operational continuity. In the water transport sector, Annex I identifies among the categories of entities concerned inland, sea and coastal passenger and freight water transport companies, port managing bodies, includ-

ing their port facilities, entities operating works and equipment within ports, and operators of Vessel Traffic Services (VTS). The actual applicability of these obligations depends on the entity's classification and on the relevant national transposition and implementation provisions. Complementarily, the European Union has recognised the equally strategic importance of physical security and critical infrastructure resilience through Directive (EU) 2022/2557, known as the Critical Entities Resilience (CER) Directive. This instrument establishes a framework aimed at strengthening the ability of critical entities to prevent, withstand, absorb and recover from events that may disrupt the provision of essential services, including those arising from physical threats, natural disasters, accidents and other emergency situations. In the transport sector, the CER framework also covers specific categories of entities operating in the maritime domain, including port managing bodies, subject to the identification criteria and procedures established by European and national legislation. The population of entities covered by CER does not fully coincide with that covered by NIS2: the two frameworks have distinct scopes of application and identification criteria, although they may apply to the same infrastructure or organisation. Their complementarity therefore makes it possible to address, from different but converging perspectives, the risks that may compromise the continuity of essential services. From this perspective, cyber protection cannot be separated from adequate physical protection, just as physical security cannot be effectively achieved without considering the digital vulnerabilities of systems and infrastructure. Unauthorised access to technical premises, network equipment, industrial control systems or operational centres may provide the initial conditions for a cyber incident. Conversely, the compromise of information systems may enable interference with physical security devices, video surveillance, access control and automation systems. For critical maritime infrastructure, characterised by strong interdependencies between physical, digital and operational components, it is therefore necessary to adopt an integrated risk management model in which physical and cybersecurity measures are coordinated, proportionate to the criticality of the infrastructure and commensurate with the identified threat scenarios. This approach should translate into a joint assessment of vulnerabilities, clearly defined responsibilities, coordinated prevention and response procedures, and periodic verification of the effectiveness of the measures implemented. Only through such integration can effective resilience be achieved, simultaneously safeguarding operational continuity, navigational safety and the availability of essential services.

An Interconnected Maritime Ecosystem and Emerging Challenges

Protecting the maritime sector therefore requires a unified view of the components that sustain its operations. Ships, ports, remote operation centres, traffic management systems, logistics platforms and communication networks form

an interdependent ecosystem in which a cyber incident affecting a single system may spread to other services and infrastructure. This complexity is further increased by the growing use of maritime drones and unmanned underwater vehicles for inspection, surveillance, research and support to maritime operations. At the same time, submarine telecommunications cables and offshore energy infrastructure, including subsea pipelines, are essential components of international connectivity and continuity of supply. These assets introduce new areas of exposure and raise challenges concerning physical protection, communications security, the control of autonomous systems and the ability to detect and manage anomalous events promptly.

Purpose and Structure of the Chapter

In light of these considerations, this publication aims to examine the main cybersecurity implications for the maritime sector, with particular reference to the implementation of the NIS2 Directive and the protection of critical infrastructure and essential services associated with navigation. The analysis will address the interdependencies between ships and ports, the security of digital platforms and traffic management systems, the operational specificities of autonomous ships and their remote operation centres, as well as the emerging challenges associated with maritime drones and subsea infrastructure. The objective is to provide a structured overview of the risks, regulatory obligations and technical, organisational and procedural measures required to strengthen the sector's cyber resilience. In a context where maritime security and digital security are increasingly interconnected, the ability to prevent, detect and manage cyber incidents is essential to ensuring the continuity of maritime traffic, the protection of infrastructure and the stability of international logistics chains.

2. The European and National Regulatory Framework for Maritime Cybersecurity: From NIS1 to NIS2 and Integration with CER

From NIS1 to NIS2: The Evolution of the European Cybersecurity Strategy

The evolution of the European cybersecurity regulatory framework reflects a growing recognition that the security of network and information systems is essential to the functioning of the internal market and the continuity of services fundamental to society. A significant milestone was Directive (EU) 2016/1148, commonly known as the NIS1 Directive, which introduced a common framework of measures designed to achieve a high level of security of network and information systems across the European Union. The Directive represented the Union's first comprehensive intervention aimed at reducing disparities between national legal frameworks, promoting cooperation among

Member States, and establishing security and incident-reporting obligations for operators of essential services and digital service providers. However, the implementation of NIS1 revealed limitations arising from fragmented national criteria for identifying covered entities, differences in security levels, and the increasing dependence of strategic sectors on interconnected digital infrastructure. The expansion of cyber threats, the growing complexity of supply chains and the possibility that a single incident may affect multiple countries and economic sectors made it necessary to strengthen the European framework. In this context, Directive (EU) 2022/2555, known as the NIS2 Directive, was adopted, repealing NIS1 and expanding the requirements relating to risk management, management-body accountability, supply chain security, and incident handling and reporting (EU, 2022).

Scope of Application of NIS2 in the Maritime Sector

The NIS2 Directive identifies transport as a critical sector and expressly includes water transport in Annex I, sector 2, subsection (c). This reflects the strategic role of maritime infrastructure and the services that support its operation, recognising that disruption of digital systems may affect traffic continuity, port call management, navigational safety and logistics chains. More specifically, the Directive covers inland, sea and coastal passenger and freight water transport companies, port managing bodies and their port facilities, entities operating facilities and equipment within ports, and operators of Vessel Traffic Services (VTS). With regard to maritime transport, the provision refers to the definitions set out in Annex I to Regulation (EC) No 725/2004 and specifies that individual vessels operated by transport companies are not, as such, included in the relevant sectoral category. Correct identification of the entities concerned nevertheless requires an assessment of the organisation's legal and operational nature, the type of service provided, its size and any applicable specific provisions. Operating in the maritime or port sector does not, by itself, automatically trigger every NIS2 obligation: the criteria established by the Directive and national implementing legislation must be assessed, together with the organisation's classification as an essential or important entity.

In this context, Italy has developed a specific initiative to strengthen cybersecurity across the maritime and port sector through Circular Navigation Safety – General Series No. 177/2025, issued by the Ministry of Infrastructure and Transport (MIT), through the Headquarters of the italian Coast Guard, in coordination with the NIS Authority for the Transport Sector. Entitled Maritime Cyber Risk: Updating Security Measures for Italian-Flagged Ships, ISM Management Companies and Port Facility Operators, the Circular represents a significant initiative in the integration of cybersecurity into maritime security procedures. It establishes specific guidance and measures for

Italian-flagged ships, management companies and port facilities, promoting a structured approach to cyber risk management and the protection of digital and operational systems. It also introduces specific roles responsible for cybersecurity: the Cyber Company Officer (CyCO), the Cyber Security Officer (CySO) and the Cyber Port Facility Security Officer (CyPFSO), with dedicated training envisaged for these roles in line with their assigned responsibilities. The initiative therefore helps translate general cyber risk management principles into operational guidance tailored to the specific characteristics of ships and port facilities. This initiative is complemented by the interpretative and guidance activities undertaken as part of the national implementation of NIS2. The Ministry of Infrastructure and Transport, acting as the NIS Authority for the transport sector, has contributed to clarifying the identification of port-based entities to which the framework applies, through FAQs made available on the official portal of the National Cybersecurity Agency (Agenzia per la cybersicurezza nazionale, ACN). These clarifications are particularly important in distinguishing port managing bodies from other economic operators located within port areas and in assessing whether NIS2 obligations apply on the basis of the activities actually performed and the relevant statutory criteria. Coordination between the MIT's sector-specific guidance and the national framework established by ACN is therefore an essential instrument for ensuring consistent implementation and clarifying responsibilities within an ecosystem characterised by public and private entities, shared services and interdependent infrastructure.

Safety Requirements, Incident Management, and Organizational Responsibilities

The core of the NIS2 framework is Article 21, which requires essential and important entities to adopt appropriate and proportionate technical, operational and organisational measures to manage risks to the security of network and information systems used in their operations or service provision. These measures must follow an all-hazards approach and include, among other elements, risk analysis, incident handling, business continuity, system recovery, supply chain security, vulnerability management, cryptography, access control and human resources security.

For maritime operators, these obligations must be translated into measures consistent with the actual characteristics of the services provided. A VTS system, a port access management platform, a ship-to-shore communications infrastructure and a terminal automation system present different risk profiles, reflecting their respective functions and the consequences of a potential compromise. Risk assessments must therefore also consider dependencies between information technology and operational technology (OT) systems, service avail-

ability, data integrity and the possibility that an incident may propagate to other entities within the maritime ecosystem.

A further central element is the accountability of management bodies, addressed in Article 20 of the Directive. Management bodies are required to approve cybersecurity risk management measures, oversee their implementation and promote the necessary training. Cybersecurity therefore becomes a governance matter: it is no longer merely a technical function entrusted to IT personnel, but an integral component of organisational decisions, investment planning, supplier management and service continuity planning. NIS2 also strengthens the reporting framework for significant incidents under Article 23. The European incident-reporting framework requires an early warning within 24 hours of becoming aware of a significant incident, an incident notification within 72 hours and, as a general rule, a final report within one month of the initial notification, subject to the conditions set out in the provision. For maritime organisations within scope, compliance with these deadlines requires effective escalation procedures, monitoring, evidence collection, impact assessment and coordination with the competent authorities and CSIRTs.

The CER Directive: From Cybersecurity to the Overall Resilience of Critical Infrastructure

Alongside the framework governing digital security, the European Union has developed a complementary framework dedicated to the resilience of critical entities. Directive (EU) 2022/2557, known as the Critical Entities Resilience (CER) Directive, repeals the previous Directive 2008/114/EC and establishes provisions designed to strengthen the ability of critical entities to prevent, withstand, absorb and recover from events that may disrupt the provision of essential services. CER addresses a broad range of risks, including intentional threats, sabotage, terrorism, natural disasters, public health emergencies and other events that may affect operational continuity. In the transport sector, the CER Annex includes the maritime subsector, including maritime ports and port facilities falling within the applicable legal definitions. Its practical application nevertheless depends on the identification of critical entities by the Member State, based on risk assessment and the criteria established by the Directive. It is therefore incorrect to assume that every organisation subject to NIS2 is automatically a critical entity under CER, or vice versa. The two regulatory frameworks may overlap at organisational and infrastructure level, but pursue specific objectives through distinct identification criteria and obligations. CER establishes, among other requirements, obligations for Member States and critical entities to conduct risk assessments, implement resilience measures proportionate to identified risks, designate relevant contact points, establish incident-handling procedures and report relevant events to the competent au-

thorities. For port infrastructure, this entails analysing threats such as unauthorised physical access, sabotage of facilities and equipment, disruption of energy and communication services, infrastructure damage and the unavailability of essential resources. Resilience must be assessed in terms of the organisation's ability to maintain or restore service provision, including during severe or concurrent events.

National Implementation: Legislative Decree No. 138/2024 and Legislative Decree No. 134/2024

In Italy, the NIS2 Directive was transposed through Legislative Decree No. 138 of 4 September 2024, which governs the identification of covered entities, cybersecurity risk management obligations, incident reporting, supervisory activities and the associated enforcement framework. The Decree assigns a central role to the National Cybersecurity Agency (Agenzia per la cybersicurezza nazionale, ACN) in implementing the national cybersecurity framework and establishes a system of authorities and responsibilities involving sectoral administrations. Its application to the maritime sector must therefore be considered in light of both the Decree's general provisions and the implementing measures and guidance issued by the competent authorities. In particular, determining the regulatory status of maritime entities requires verification of their inclusion within the scope established by national legislation, their classification as essential or important entities, and any specific obligations arising from the nature of their services. For organisations concerned, compliance cannot be achieved merely by deploying technological protection solutions. It requires a documented risk management system, clearly assigned responsibilities, incident management procedures, verification of implemented controls and the ability to demonstrate effective fulfilment of applicable obligations. In parallel, the CER Directive was transposed through Legislative Decree No. 134 of 4 September 2024, which governs the identification of critical entities, risk assessment, resilience measures, supervision and institutional coordination. The Decree establishes a national framework focused on the continuity of essential services and the ability of designated organisations to withstand events of different natures. Here too, operating in the maritime sector does not automatically confer critical-entity status: the national identification procedures and the criteria established under the CER framework must be applied.

Integration between NIS2 and CER: A Unified Approach to the Protection of Maritime Infrastructure

The relationship between NIS2 and CER is particularly relevant in the maritime sector, where physical and digital components are closely interdependent. Port operations, vessel traffic management, communications systems,

automation devices and energy infrastructure contribute to the provision of essential services and may be affected by threats that manifest simultaneously in the physical and cyber domains. Article 1(2) of the CER Directive requires Member States to ensure coordinated implementation of CER and NIS2, specifically in view of the relationship between physical security and cybersecurity. A cyberattack against port management systems may disrupt port call planning, terminal access and cargo handling. Similarly, physical sabotage may interrupt power supplies, damage communications equipment or render essential information systems unavailable. In other scenarios, the compromise of a privileged account or remote access platform may facilitate interference with control and automation devices. These interdependencies make it necessary to move beyond a rigid separation between physical and digital security and adopt an integrated view of risk scenarios and their potential operational consequences. Operational integration should begin with a coordinated assessment of threats and vulnerabilities, linking physical assets to the IT and OT systems on which they depend. This assessment should identify essential functions, interconnection points, dependencies on external suppliers and the potential cascading effects of disruption. On this basis, measures proportionate to asset criticality can be established, including physical and logical access controls, network segmentation, event monitoring, communications protection, service redundancy, recovery procedures and coordinated emergency response plans. Convergence between the two regulatory frameworks does not mean that their respective obligations are identical or automatically interchangeable. NIS2 remains focused on cybersecurity risk management and the security of network and information systems, whereas CER focuses on the resilience of critical entities against the broader range of threats that may disrupt essential services. Where the same organisation falls within both frameworks, risk assessments, responsibilities, continuity plans and incident management procedures should be harmonised while preserving the specific requirements and obligations of each regime.

Implications for Maritime Cybersecurity Governance

Effective implementation of NIS2 and CER in the maritime sector requires a governance model that connects regulatory obligations with the day-to-day management of infrastructure. This entails coordinated involvement of cybersecurity, physical security, network management, automation systems, business continuity and emergency management functions. Responsibilities must be clearly assigned, supported by communication and escalation procedures that enable relevant information to be shared promptly among technical teams, organisational management and competent authorities. A further consideration concerns interdependencies between organisations. Ports, shipping companies, terminal operators, VTS providers, connectivity suppliers and energy infrastructure operators

may rely on shared systems and services. The resilience of an individual entity is therefore insufficient unless accompanied by an understanding of external dependencies and procedures for handling incidents affecting multiple operators. Supply chain security and cooperation between public and private entities are consequently essential to reducing systemic risk. In conclusion, the transition from NIS1 to NIS2 and the introduction of the CER framework mark a shift towards a concept of maritime security based on the overall resilience of infrastructure and services. Cyber protection and physical protection should be treated as complementary dimensions of a single risk management system, with measures proportionate to infrastructure criticality, threat scenarios and the potential consequences of disruption for society and the economy. This approach provides the regulatory and organisational foundation for addressing the challenges posed by port digitalisation, the growing connectivity of ships, and the development of autonomous systems and distributed maritime infrastructure.

3. Maritime Drones: Cybersecurity of Surface and Underwater Unmanned Systems

The Evolution of Unmanned Systems in the Maritime Domain

Technological developments in the maritime sector are driving the progressive adoption of unmanned systems designed to complement the operational capabilities of maritime authorities, surveillance organisations and port infrastructure operators. The term maritime drone encompasses technologies that differ in their operating environment, autonomy, propulsion and control methods: from Remotely Piloted Aircraft Systems (RPAS), used for aerial surveillance, to Unmanned Surface Vehicles (USVs), and autonomous or remotely operated underwater vehicles, including Autonomous Underwater Vehicles (AUVs) and Remotely Operated Vehicles (ROVs). These systems are not necessarily alternatives to conventional vessels; rather, they are complementary assets that extend personnel's operational reach, reduce exposure to hazards and collect information in areas that are difficult to access. Their effectiveness increasingly depends on their integration into a coordinated operational system in which different platforms cooperate through communications networks, sensors and command-and-control centres.

Surface Drones: Surveillance, Patrol, and Maritime Security

Unmanned Surface Vehicles (USVs) represent an emerging component of maritime surveillance. They can support patrol operations, environmental data collection, port area monitoring, infrastructure observation and search-and-rescue (SAR) activities. Depending on their configuration, they may be

remotely controlled, follow predefined routes or use increasing levels of autonomy for navigation and mission execution. Their deployment can extend operational presence and collect information without directly exposing personnel to certain hazardous conditions. Increasing autonomy does not remove the need to secure command-and-control systems. Communications between the operator and the vehicle, satellite navigation systems, sensors, onboard software and mission-planning platforms must all be included in the risk assessment. Interference with positioning, loss of the remote link, manipulation of sensor data or unauthorised access to control systems may compromise mission reliability and, under certain circumstances, create risks for other vessels or infrastructure. Security must therefore be built in from the outset, addressing authentication, authorisation, communications protection, secure software updates, event logging and recovery procedures in the event of loss of control.

Underwater Drones: Inspection, Search, and Infrastructure Protection

Unmanned underwater systems further expand maritime authorities' operational capabilities, enabling activities that would be difficult, costly or hazardous to perform through conventional diving operations. ROVs, connected to their control unit by an umbilical cable, enable inspections and real-time data transmission, whereas AUVs can carry out autonomous missions based on predefined navigation plans and data acquisition profiles. Underwater services can support maritime accident investigations, the recovery of voyage data recorders (VDRs), the search for submerged objects, three-dimensional surveys and the monitoring of maritime infrastructure. These technologies are particularly relevant to the protection of ports, submarine cables, pipelines, anchoring systems and other critical infrastructure located on the seabed. Sonar, cameras, magnetometers and other sensors can help reconstruct the underwater environment, identify anomalies and document the condition of structures, for example the work-class ROV services supported by sonar surveys, live data transmission and three-dimensional modelling. These tools can improve inspection and response capabilities, but their effectiveness depends on measurement quality, correct interpretation of findings and integration with other available information sources.

Interconnection Between Aerial, Surface, and Underwater Systems

The operational value of maritime drones increases when aerial, surface and underwater capabilities are coordinated within a single mission architecture. An RPAS (Remotely Piloted Aircraft System) may provide an overview of the affected area, locate a vessel or a person in distress, and guide subsequent operations by a USV (Unmanned Surface Vehicle) or ROV (Remotely Operated Vehicle) or UUV (Unmanned Underwater Vehicle). A surface vehicle may collect envi-

ronmental information or act as a support platform, while an underwater system may conduct a more detailed inspection of a wreck, the seabed or infrastructure. The objective is not to add independent platforms, but to create an interoperable capability in which different systems contribute to a common operational picture. Operational initiatives supported by EFCA and EMSA in the EU demonstrate how aerial drones can complement vessels in monitoring and surveillance activities, while underwater systems can provide information that cannot be obtained through surface observation alone. Integrating these data with satellite information, maritime traffic monitoring systems and geospatial analysis tools can broaden the information available to authorities. Also, Frontex on the European Action Plan for drone and counter-drone security also emphasises interoperability, common data formats, integration into national surveillance systems and the sharing of the operational picture at European level.

The Cyberattack Surface of Marine Drones

The introduction of drones into the maritime domain expands operational capabilities but also increases the cyberattack surface. Each platform incorporates hardware, software, communications systems, sensors, navigation equipment and command interfaces that may constitute potential vulnerabilities. The risk concerns not only the vehicle itself but also control stations, data transmission networks, remote processing services, storage platforms and connections to the organisation's information systems. An attack may seek to disrupt a mission, alter collected information, compromise data confidentiality or induce an operator to make decisions based on an inaccurate representation of the situation. Relevant threat scenarios include unauthorised access to control stations, credential compromise, interference with communications links, manipulation of navigation data and compromise of software or updates. For underwater vehicles, risks relating to the availability of physical or radio links, the integrity of collected data and the correct synchronisation of information must also be assessed. In autonomous systems, the possibility of continuing a mission without continuous human control makes it particularly important to verify the reliability of automated functions, the limits of autonomy and the expected behaviour in the event of anomalies. These scenarios represent a risk assessment based on system characteristics and do not imply that every platform necessarily exhibits all the vulnerabilities described. Protection must therefore cover the platform's entire lifecycle: supplier selection and qualification, initial configuration, identity and access management, communications security, software updates, event monitoring, log retention and secure decommissioning. Dependencies on third-party components and services must also be assessed, responsibilities for vulnerability management must be defined, and incident response procedures must include vehicle recovery and verification of data integrity. For organisa-

tions subject o the NIS2 Directive, these measures must be incorporated into the organisation's risk management framework and proportionate to the criticality of the service and the consequences of a potential compromise.

Use of Drones in Maritime Search and Rescue Operations

Maritime search-and-rescue operations are among the areas in which the integrated use of drones can provide significant operational benefits. Aerial platforms can help locate people in the water, survey large areas, assess environmental conditions and update the tactical picture. Surface systems can support area monitoring and collect additional information, while underwater vehicles may be deployed, where appropriate, for subsequent search and inspection activities following an incident. Technological support must not, however, be confused with the ability to make operational decisions autonomously. During a rescue operation, data from drones must be interpreted in conjunction with information collected by vessels, coordination centres and other participating organisations. A thermal image, a georeferenced position or an automated alert constitutes an item of information that requires verification, contextualisation and professional assessment. The quality of the outcome therefore depends not only on sensor accuracy but also on the ability to integrate information, recognise its limitations and transform it into operationally useful guidance.

The Human Factor, Information Overload, and the Transformation of Command and Control Center

A central issue in the use of maritime drones concerns the relationship between technological capability and the human factor. The technology can extend operators' capabilities, but it is the human being who gives meaning to the information and determines how it is used operationally. This principle is particularly important in maritime rescue operations, where urgency, uncertainty and the need to coordinate multiple units require timely and informed decisions. The growing availability of sensor readings, images, tracks, environmental measurements, automated alerts and video streams can improve situational awareness, but it can also create information overload. During an emergency, an operator required to examine information simultaneously from multiple drones, radar systems, communications platforms, mapping tools and units on scene may struggle to distinguish priority information from secondary details. More information does not necessarily lead to a proportionate improvement in decision-making quality: if data are not filtered, correlated and presented in a manner consistent with operational needs, they may increase cognitive workload, decision-making stress and the risk of overlooking relevant signals. System design must therefore go beyond the technical integration of platforms and take into account cognitive processes, workload and the way information

is presented to operators. Command-and-control centres must be redesigned to manage an information environment in which conventional systems and new unmanned platforms operate simultaneously. This requires integrated interfaces, a common operational picture, tools for data correlation and prioritisation, escalation procedures, and mechanisms to distinguish verified information from automated alerts that still require confirmation. The objective is not to provide operators with an ever-increasing volume of data, but to deliver the information they actually need, at the right time and at an appropriate level of detail for the decision to be made.

Artificial Intelligence, Decision Support, and Human Responsibility

Artificial intelligence can make a significant contribution to managing the information complexity associated with drone operations. Automated analysis tools can support image classification, object detection, track correlation, anomaly detection and alert prioritisation. In a SAR context, these capabilities may help identify potential persons in the water, highlight areas requiring further verification and reduce the time needed to examine large volumes of data. Similarly, in underwater inspections, automated processing of imagery and sonar data can facilitate three-dimensional reconstruction and comparisons between successive surveys. These tools must nevertheless be assessed in terms of output reliability, data quality, false positives and false negatives, and the possibility that operational conditions differing from those encountered during development may reduce their performance. Artificial intelligence should therefore be conceived as a decision-support tool rather than an automatic substitute for professional judgement. A system can process data, identify correlations, propose scenarios and flag priorities, but the final operational decision must remain with competent human personnel who possess the necessary authority and understand the limitations of the tools being used. This principle is essential in rescue operations, where decisions may directly affect human life, but it is equally relevant to surveillance, maritime security and the protection of critical infrastructure. Decision-making accountability must be supported by transparency, traceability of automated recommendations, the ability to verify results and procedures for intervening when a system produces unexpected or insufficiently reliable outputs.

Staff Training and Operational Capacity Building

The availability of technologically advanced systems does not, by itself, guarantee improved operational performance. Personnel must be trained not only to operate individual platforms but also to understand the data they produce, assess information reliability, manage anomalies and integrate findings into the decision-making process. Training should cover operational, digital and cybersecurity

skills, together with exercises simulating complex scenarios involving multiple information sources, degraded communications, conflicting reports and limited decision time. An effective model should also foster cooperation among drone operators, command-centre personnel, IT specialists, security managers and maritime operations coordinators. Such cooperation enables common procedures to be established, anomalies to be identified promptly, assessment criteria to be shared and the risk of relevant information remaining isolated in non-interoperable systems to be reduced. Preparedness should be periodically tested through joint exercises and post-operation reviews, allowing procedures to be updated in response to identified weaknesses. Technology must therefore be accompanied by sound organisation, continuous training and clearly assigned responsibilities.

Concluding Remarks: Technological Capability at the Service of Humanity

Aerial, surface and underwater drones are becoming an increasingly important component of evolving maritime capabilities. Their use can improve surveillance, extend intervention capacity, reduce personnel exposure to hazards and make previously difficult-to-obtain information available. These benefits, however, depend on the ability to integrate platforms, secure digital systems, ensure interoperability and transform data into information that can be used in operational processes. Growing technological complexity therefore requires an approach that jointly addresses cybersecurity, technical reliability, organisation, training and the human factor. Technology can extend operational reach, but it does not replace the human role in interpreting information and conducting the mission. During a maritime rescue, the simultaneous availability of data from multiple drones and surveillance systems can provide a decisive advantage, but only if the information is organised, filtered and presented in a way that supports, rather than obstructs, decision-making. Personnel training and the redesign of command-and-control centres are therefore essential to managing information overload and preserving situational awareness. Artificial intelligence can accelerate analysis and support decisions, but responsibility for the final operational decision must remain the responsibility of human operators. Ultimately, the value of drones lies not only in their technical performance but also in an organisation's ability to integrate them into a secure, interoperable and human-centred system.

4. Cybersecurity of Maritime Autonomous Surface Ships (MASS)

Autonomous Ships: Technological Evolution and New Cybersecurity Challenges

Maritime Autonomous Surface Ships (MASS) represent one of the most significant developments in maritime navigation, introducing increasing levels

of automation and autonomy in ship operation, navigation management and interaction with port infrastructure. Their operation depends on the integration of navigation systems, sensors, radar, AIS, ship-to-shore communications, data-processing platforms and Remote Operations Centres (ROCs). This transformation expands operational capabilities but also makes cybersecurity an essential requirement for ensuring navigational safety, operational continuity and the protection of maritime infrastructure. The introduction of MASS must be addressed through an integrated approach encompassing physical security, IT and OT systems, communications networks and port digital infrastructure, following the principle of security by design.

Regulatory Framework and International Guidelines

In addition to the aforementioned European NIS2 Directive (EU 2022/2555) and CER Directive (EU 2022/2557), a specific reference for maritime cybersecurity is IMO Circular MSC-FAL.1/Circ.3/Rev.3, containing the Guidelines on Maritime Cyber Risk Management. These guidelines promote an approach to cyber risk management integrated into maritime safety and security processes, covering ship navigation, communications, control and management systems, as well as their interconnections with port infrastructure. For Maritime Autonomous Surface Ships (MASS), these principles are particularly relevant due to their reliance on digital systems, ship-to-shore communications and Remote Operations Centres (ROCs). At the international level, the IMO is also developing a dedicated regulatory framework for MASS, aimed at harmonising the safety requirements and operational responsibilities associated with autonomous navigation.

Key Threats and Protective Measures

The absence of a crew physically present on board does not reduce the importance of security; rather, it shifts some critical functions towards digital infrastructure and ship-to-shore connections. Relevant threat scenarios include the interception or manipulation of communications, GNSS spoofing, false data injection into sensors, compromise of navigation systems, and theft or alteration of control-centre credentials. There text also highlights the need to protect ECDIS (Electronic Chart Display and Information System) systems, artificial intelligence-based applications and interfaces between ships, ROCs and port infrastructure. Countermeasures include dedicated networks, secure communications, authentication of information sources, zero-trust architectures, anomaly monitoring and business continuity procedures. In this context, IEC 62443, covering the security of industrial automation and control systems, and IEC 63173-2, concerning secure ship-to-shore communications through SECOM, are cited in the presentation as relevant standards for developing a resilient architecture. Protection must also extend to Vessel Traffic Services

(VTS), private 5G networks, cloud infrastructure and electronic navigation services that contribute to mission management.

The Italian 5G MASS Project in Livorno: A Model of Excellence and Innovation

A particularly significant example is the 5G MASS project, developed within the research ecosystem of the Port of Livorno by the National Inter-University Consortium for Telecommunications (CNIT), with the involvement of the Italian Coast Guard and other project partners. The CNIT laboratory in Livorno provides an advanced test environment for developing and validating digital solutions for navigation and port infrastructure. The activities coordinated by Prof. Paolo Pagano include trials involving autonomous ships, 5G networks, VTS systems, satellite communications, electronic navigation services and remote operations centres. The project's results have also been shared within the IMO context, as documented by reference MSC.110/INF.5. This initiative represents an Italian example of excellence and a pioneering approach in the field, enabling the interaction between autonomous ships, port infrastructure and supervision systems to be tested in a real operational environment. The model has the potential to be replicated in other ports and countries, providing a technological and methodological foundation for developing shared standards, procedures and practices. Its value lies not only in testing autonomous navigation but also in generating scientific evidence to support international discussions and the evolution of the IMO regulatory framework.

Regulatory Perspectives and Conclusions

The Livorno project is particularly important during the ongoing development of the international regulatory framework for MASS. The IMO's voluntary code through Resolution MSC.595(111), which entered into effect on 1 July 2026, is the subsequent process towards a mandatory code, which, according to the current schedule, is expected to enter into force on 1 January 2032. The Experience-Building Phase provides an opportunity to collect trial results, assess risks, verify the effectiveness of security measures and contribute to defining requirements applicable to ships, ROCs and communications infrastructure. Within this process, Italian experience can make a significant scientific and operational contribution by promoting the international transferability of tested solutions and fostering dialogue between researchers, maritime authorities and regulatory bodies. Cybersecurity must be treated as a fundamental requirement of autonomous navigation, incorporated from the design stage and verified throughout the system lifecycle. Only by combining experimentation, standardisation, digital resilience and human oversight will it be possible to support the safe and sustainable development of MASS.

5. Cybersecurity of Submarine Infrastructure: Cables, Pipelines and Underwater Operations

Strategic Relevance and Interdependence of Submarine Infrastructure

Submarine infrastructure is a key component of States' economic, technological and strategic security. Submarine telecommunications cables support a substantial share of global digital communications, while pipelines enable the transportation of energy resources. Damage to these assets can have consequences far beyond the physical location of the incident, affecting service continuity, economic activity and national security. Their protection therefore requires an approach that jointly considers the infrastructure itself, vessels, monitoring systems, maritime activities and the operators involved in installation, maintenance and repair. Submarine infrastructure cannot be effectively protected through isolated or exclusively emergency-driven measures. Its resilience depends on the interactions between the seabed, navigation and fishing activities, specialised vessels, command-and-control systems and communication networks. Infrastructure security must therefore be treated as the outcome of a continuous process encompassing prevention, monitoring, anomaly detection, response and recovery, embedded within the wider maritime security system.

Physical, Cyber and Hybrid Threats

Threats to submarine infrastructure include accidental events, natural phenomena and deliberate actions. Anchors, fishing gear, vessel operations, wear and environmental conditions can cause damage even in the absence of hostile intent. At the same time, apparently lawful maritime activities may make deliberate interference difficult to identify promptly. Risk assessment must therefore distinguish between the cause of an incident, intent, impact and the possibility of attributing responsibility, without automatically treating every instance of damage as a hostile act. The increasing digitalisation of underwater operations introduces an additional layer of exposure. Mission command-and-control systems, Dynamic Positioning (DP), navigation and communication systems, and ship-to-shore interfaces may be interdependent. The compromise or unavailability of one component can affect others, with consequences for vessel control, underwater vehicle management and operational continuity. Cyber risk therefore acquires a cyber-physical dimension, in which a digital anomaly can produce tangible consequences for navigational safety, the marine environment and infrastructure integrity. Security must also cover operational and mission-related information. Data concerning vessel positions, routes, operational time windows, the characteristics of deployed systems and infrastructure status may be highly sensitive. Their alteration, disclosure or unavailability can compromise the safety of an intervention, facilitate unauthorised interference or

hinder service restoration. Protection must therefore extend beyond technological systems to include information flows and interfaces between the various stakeholders involved.

Cybersecurity of Underwater Operations and Specialised Vessels

Cable-laying vessels and ships involved in the installation, inspection, maintenance, repair and recovery of submarine infrastructure are essential to overall resilience. Their operations rely on the integration of onboard systems, positioning equipment, communication networks and remotely operated vehicles (ROVs) or autonomous underwater vehicles (AUVs). Cybersecurity must therefore be assessed across the entire operational ecosystem and its interdependencies, rather than focusing exclusively on individual devices. Particular attention should be paid to Dynamic Positioning systems, mission command-and-control systems, and communications between the vessel, underwater vehicle and shore-based support centre. These elements should be included in risk assessments, with particular regard to data availability, integrity and reliability. Temporary connections, digital interfaces and the presence of external personnel must also be considered, ensuring that operational integration does not introduce unauthorised access or compromise pathways between systems with different functions. Operational continuity and recovery must be integral to security planning. In the underwater environment, restoring functionality may be complex and cannot always be achieved immediately, owing to vehicle location, environmental conditions and the stage of the operation. Incident-response procedures must therefore reflect navigational safety and environmental requirements, providing for anomaly management, controlled recovery and the safe resumption of activities. The objective is not merely to prevent cyber incidents, but also to limit their effects on the entire operation and the infrastructure concerned.

Practical Example of an Integrated Regulatory Approach to Submarine Infrastructure Protection

Consider, by way of example, a submarine telecommunications cable connecting two European countries and operated by an entity falling within the scope of the critical infrastructure regulatory framework. Protecting this infrastructure provides a practical illustration of how European and national provisions on cybersecurity, resilience and maritime security can be implemented in a coordinated manner. This hypothetical case shows how regulatory requirements can be translated into operational measures throughout the infrastructure lifecycle: from risk assessment and monitoring to incident management and service recovery. In the first phase, the operator conducts an integrated risk assessment, identifying physical and cyber threats that could compromise service availability. The analysis includes, among other factors, accidental damage caused by anchors or fishing gear, natural events,

deliberate interference and compromise of the digital systems used to monitor and manage the infrastructure. At this stage, the CER Directive and its national implementation provide the framework for assessing the resilience of the critical entity, while the NIS2 Directive provides the reference for cyber risk management, insofar as the operator and the services concerned fall within their respective scopes. At the organisational level, the operator establishes responsibilities, escalation procedures and communication channels with the competent authorities and other stakeholders across the supply chain. It also prepares a business continuity and recovery plan identifying intervention priorities, available resources, coordination arrangements with specialised vessels and procedures for restoring services following an outage. Measures must be proportionate to the identified risks and consistent with the obligations actually applicable, avoiding unnecessary duplication across different regulatory frameworks. The cyber component is addressed by protecting the information and operational systems supporting cable monitoring, alarm management, communications and maintenance coordination. Measures may include controlled access management, network segmentation where appropriate, system updates, protection of remote interfaces, log collection and analysis, and periodic testing of incident-response procedures. The objective is to prevent the compromise of a supporting system from spreading to other essential components or delaying the detection of physical damage. Suppose the monitoring system reports an anomaly while maritime traffic data indicate the presence of a vessel near the affected location. The operator activates verification procedures and shares relevant information with the competent authorities, without automatically assuming malicious intent. A joint assessment of the available evidence helps distinguish, insofar as possible, between a technical fault, accidental damage and possible deliberate interference, guiding a proportionate response consistent with the applicable legal framework. Where physical intervention on the cable is required, inspection and repair activities also take account of the cybersecurity of the specialised vessels involved. Dynamic Positioning systems, ship-to-shore communications and underwater vehicle control systems must be considered as part of the operational risk assessment. In this context, for example, the Circular No. 177/2025 issued by the Italian Coast Guard provides the national general reference for maritime cyber risk management; specific guidance on underwater operations and cable installation and maintenance can complement this framework once the final details and formal status of the relevant instrument have been verified. For informational purposes, the Italian Coast Guard is currently working on issuing a circular similar to the one mentioned, but applicable to the underwater domain. This example demonstrates that the coordinated implementation of NIS2 and CER does not simply involve adding together documentary compliance requirements. Rather, it creates a unified process in which risk assessment informs preventive measures, monitoring supports early detection, institutional cooperation enables a coordinated response, and continuity

and recovery plans limit the effects of disruption. Infrastructure resilience therefore results from the integration of governance, physical protection, cybersecurity and recovery capabilities, while respecting the responsibilities and obligations applicable to each entity.

An integrated maritime security model

The resilience of submarine infrastructure cannot rely exclusively on seabed surveillance or defensive measures activated after an incident. A security model based on information sharing, maritime traffic monitoring, the detection of anomalous behaviour and cooperation between competent authorities, infrastructure operators, shipping companies and technology providers is required. Existing surface maritime-monitoring capabilities can contribute to situational awareness in the underwater domain, although they cannot replace the specialised technologies and expertise required for the direct inspection and protection of submarine infrastructure. Responses to threats must also be based on a proportionate and documented assessment that distinguishes accidental incidents from deliberate conduct and takes the available evidence into account before attributing responsibility. Maritime security measures, data collection and validation, and institutional coordination can support graduated responses consistent with the applicable legal framework, reserving more intrusive measures for situations in which the relevant conditions are met. Resilience and defence are not mutually exclusive concepts: they are complementary instruments to be applied according to the nature of the threat, the operational context and the relevant legal framework.

6. Conclusions

Submarine infrastructure must be considered an integral part of the maritime system, alongside ships, ports and the digital networks that support economic and institutional activities. Its distinctive environmental and technological characteristics require specialised expertise and measures, but do not justify separating it from the broader maritime security framework. On the contrary, the interdependencies between the surface and seabed, physical and digital systems, and public and private stakeholders make an integrated approach essential. Building a genuinely resilient system therefore requires the integration of cyber risk management, physical protection, operational continuity, recovery capabilities and institutional cooperation. Within this model, cybersecurity is neither an isolated requirement nor the exclusive responsibility of IT departments: it is a structural component of navigational safety and critical infrastructure protection. Only a coordinated maritime ecosystem, with information-sharing arrangements and clearly defined responsibilities, can effectively address complex threats and incidents with potentially systemic consequences.

References

Commission Recommendation (EU) 2024/779 of 26 February 2024 on Secure and Resilient Submarine Cable Infrastructures.

EMSA Underwater Services – https://emsa.europa.eu/we-do/safety/usw.html .

European Commission – Action Plan on Drone and Counter Drone Security – 2026.

European Commission – EU Drone Sector state of play – A Drone Strategy 2.0 -2022.

European Commission – Joint communication to the European parliament and the council – EU Action Plan on Cable Security 21 February 2025.

European Commission – The European Ocean Pact 2025, specifically Article 6.1.

European Parliament and Council of the European Union. (2016). Directive (EU) 2016/1148 of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union. Official Journal of the European Union, L 194, pp. 1–30. Official source: EUR-Lex – Directive (EU) 2016/1148 (NIS1).

European Parliament and Council of the European Union. (2016). Directive (EU) 2016/1148 of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union. Official Journal of the European Union, L 194, 19 July 2016, pp. 1–30. Official text: EUR-Lex – Directive (EU) 2016/1148.

European Parliament and Council of the European Union. (2019). Regulation (EU) 2019/1239 of 20 June 2019 establishing a European Maritime Single Window environment and repealing Directive 2010/65/EU. Official Journal of the European Union, L 198, pp. 64–87.

European Parliament and Council of the European Union. (2022). Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing --Directive (EU) 2016/1148 (NIS2 Directive). Official Journal of the European Union, L 333, pp. 80–152. Official source: EUR-Lex – Directive (EU) 2022/2555 (NIS2).

European Parliament and Council of the European Union. (2022). Directive (EU) 2022/2557 of 14 December 2022 on the resilience of critical entities and repealing Council Directive 2008/114/EC (CER Directive). Official Journal of the European Union, L 333, pp. 164–198. Official source: EUR-Lex – Directive (EU) 2022/2557 (CER).

European Parliament and Council of the European Union. (2022). Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS2 Directive). Official Journal of the European Union, L 333, 27 December 2022, pp. 80–152. In particular, Articles 20, 21 and 23 and Annex I, sector 2, subsection (c). Official text: EUR-Lex – Directive (EU) 2022/2555.

European Parliament and Council of the European Union. (2022). Directive (EU) 2022/2557 of 14 December 2022 on the resilience of critical entities (CER Directive). Official Journal of the European Union, L 333, 27 December 2022, pp. 164–198. In particular, Article 1(2), Articles 5–13 and the Annex concerning the transport sector. Official text: EUR-Lex – Directive (EU) 2022/2557.

European Union. Directive (EU) 2022/2555 (NIS2).

European Union. Directive (EU) 2022/2557 (CER) on the resilience of critical entities.
IEC. IEC 62443 series: Security for industrial automation and control systems.
IEC. IEC 63173-2: Maritime navigation and radiocommunication equipment and systems – Data interfaces – Part 2: Secure communication (SECOM).
IMO. MSC-FAL.1/Circ.3/Rev.3, Guidelines on Maritime Cyber Risk Management.
IMO. MSC.1/Circ.1639, Guidelines on Cyber Security Onboard Ships.
IMO. MSC.110/INF.5, document referenced in the 5G MASS project presentation.
IMO. Resolution MSC.595(111), MASS Code, as referenced in the 2026 presentation.
International Maritime Organization (IMO). (2026). IMO Adopts First Global Code for Autonomous Ships. Press briefing, 22 May 2026. Official source: IMO – International Code of Safety for MASS.
Italian Ministry of Infrastructure and Transport (MIT), Headquarters of the italian Coast Guard. (2025). Navigation Safety – General Series Circular No. 177/2025: Maritime Cyber Risk. Updating Security Measures for Italian-Flagged Ships, ISM Management Companies and Port Facility Operators. Issued on 16 December 2025.
Italian National Cybersecurity Agency (ACN). NIS2 Directive – FAQs and guidance for entities within the scope of the Italian national cybersecurity framework. Official ACN portal. The specific MIT guidance concerning port operators should be cited by its exact title and publication date once the relevant FAQ page or document has been identified. Official portal: Italian National Cybersecurity Agency (ACN)
Italian Republic. (2024). Legislative Decree No. 134 of 4 September 2024 implementing Directive (EU) 2022/2557 on the resilience of critical entities and repealing Council Directive 2008/114/EC. Official Gazette of the Italian Republic, General Series No. 223, 23 September 2024. Official source: Italian Official Gazette – Legislative Decree No. 134/2024.
Italian Republic. (2024). Legislative Decree No. 134 of 4 September 2024 implementing Directive (EU) 2022/2557 on the resilience of critical entities. Official Gazette of the Italian Republic, General Series No. 223, 23 September 2024. Official text: Normattiva – Legislative Decree No. 134/2024.
Italian Republic. (2024). Legislative Decree No. 138 of 4 September 2024 implementing Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union. Official Gazette of the Italian Republic, General Series No. 230, 1 October 2024.
Italian Republic. (2024). Legislative Decree No. 138 of 4 September 2024 implementing Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union. Official Gazette of the Italian Republic, General Series No. 230, 1 October 2024. Official text: Gazzetta Ufficiale – Legislative Decree No. 138/2024.
Italy. Legislative Decree No. 138 of 4 September 2024.
Lab-Ports – CNIT https://jlab-ports.cnit.it or 5G MASS | ESA Space Solutions, https://business.esa.int/projects/5g-mass
Official source: EUR-Lex – Regulation (EU) 2019/1239. European Commission. European Maritime Single Window environment (EMSWe). Official information on the European framework for harmonised ship reporting formalities. Official source: European Commission – EMSWe.

Official source: Italian Official Gazette – Legislative Decree No. 138/2024.
Reg. (CE) n. 768/2005 – EFCA.
Reg. (EC) No 1406/2002 EMSA Founding regulation.
Source: www.guardiacostiera.gov.it
United Nations Conference on Trade and Development (UNCTAD). (2025). Review of Maritime Transport 2025: Staying the Course in Turbulent Waters. United Nations, Geneva, 24 September 2025. Official source: UNCTAD – Review of Maritime Transport 2025.
United Nations Conference on Trade and Development (UNCTAD). (2026). Strait of Hormuz Disruptions: Implications for Global Trade and Development. UNCTAD/OSG/TT/INF/2026/1, 10 March 2026. Official source: UNCTAD – Strait of Hormuz Disruptions.
United Nations Convention on the Law of the Sea (UNCLOS) 10 December 1982 – specifically Articles n. 58, 79, 87, 113.

TELECOMMUNICATIONS: SECURING THE BACKBONE OF THE DIGITAL SOCIETY AND NATIONAL RESILIENCE

Matteo Macina
Telecom Italia, Italy

Abstract
Telecommunications have evolved from a connectivity utility into a strategic infrastructure on which the economy, public administration, healthcare, energy, defense and emergency services depend. Their security has therefore become inseparable from national resilience. This chapter examines how the transition towards 6G, the convergence with cloud and edge computing, the growing reliance on submarine cables and satellites, AI-driven autonomous networks and quantum computing are reshaping both the architecture and the threat landscape of digital connectivity. It analyses state-sponsored operations, cybercrime, hybrid threats and cyber warfare, and discusses NIS2 and recent European initiatives on cable security, post-quantum cryptography, cloud sovereignty and the supply chain. The chapter argues that telecommunications security must be approached as a multi-layer ecosystem challenge, and that its future will depend less on preventing every attack than on the ability to anticipate, withstand, recover and adapt.

1. Introduction: When Connectivity Becomes Strategic Infrastructure

Telecommunications are among the least visible yet most fundamental infrastructures of modern society. Every mobile call, digital payment, cloud application, emergency communication and government service rely on them. The user experience is increasingly simple. The underlying infrastructure is not. This creates a fundamental strategic reality: *connectivity has become infrastructure.*

The consequences of disruption grow accordingly. On 8 November 2023, a routing failure disconnected the Australian operator Optus for most of a day, affecting around ten million customers and disrupting access to emergency calls (Bean, 2024). On 28 April 2025, the collapse of the Iberian power grid exposed the reverse dependency: as backup batteries at mobile sites ran out, the share of users in Spain with a consistent mobile connection fell from over 90 per cent to around 40 per cent within hours (Le Maistre, 2025). Neither event was a cyberattack. Both showed how quickly a telecommunications failure becomes a societal one.

As networks become more software-defined and the geopolitical environment more contested, the sector sits at the intersection of technology, economics and national security.

2. From Networks to a Multi-Layer Digital Infrastructure

The traditional model was simple: an operator owned physical infrastructure and provided connectivity. A modern service instead depends on radio access, core and transport networks, data centres, cloud and edge platforms, submarine cables, satellites, software and APIs, identity and cryptographic infrastructure, and third-party providers (Figure 1). Each layer introduces dependencies. Each dependency introduces potential failure modes.

Telecommunications cybersecurity can therefore no longer be approached as the protection of a single network, but as the protection of an ecosystem. An operator may have mature controls around its core network and still be exposed through a supplier, a cloud service or a cable landing station. Incident data confirm it: of the 188 major telecommunications incidents reported in the EU and EFTA countries for 2024, more than a third originated in failures at third parties (European Union Agency for Cybersecurity [ENISA], 2025b).

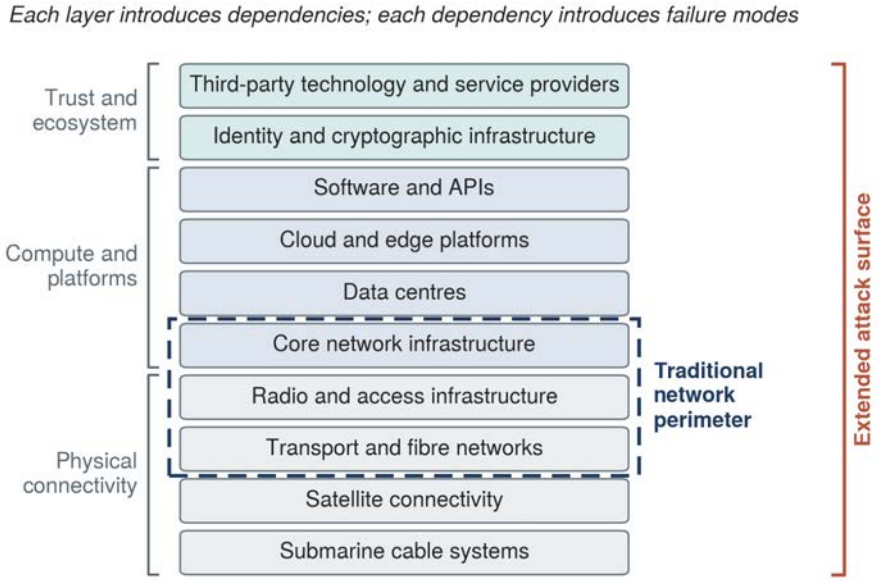

Fig. 1: The multi-layer telecommunications ecosystem and the extension of the attack surface

3. The Evolving Threat Landscape

The European threat landscape is characterized by convergence. ENISA analysed 4,875 incidents between July 2024 and June 2025: distributed deni-

al-of-service (DDoS) attacks, mostly by hacktivists, accounted for 77 per cent, ransomware remained the most impactful threat, and phishing was the main initial access vector (ENISA, 2025a).

State-sponsored actors are attracted by the intelligence, persistence and disruption potential that telecommunications offer. A joint advisory co-signed by agencies from thirteen countries described a long-running campaign, associated with Salt Typhoon, in which backbone and edge routers worldwide were modified to maintain long-term access (Cybersecurity and Infrastructure Security Agency [CISA] et al., 2025). US agencies have also assessed that Volt Typhoon had pre-positioned itself in communications and other critical networks, in some cases for at least five years, to enable disruption in a crisis (CISA et al., 2024). Defenders must therefore detect not only attacks but also *strategic positioning*.

Cybercriminals seek financial gain through ransomware, credential theft and identity-based fraud. *DDoS campaigns* place operators on both sides of the problem: they must protect their own services while their networks carry attacks against customers. *Hybrid threats* combine cyber operations with disinformation, sabotage and economic pressure below the threshold of open conflict (Giannopoulos et al., 2021).

Outage data offer an important corrective. Of the major incidents reported for 2024, system failures accounted for 60 per cent and malicious actions for only 8 per cent (ENISA, 2025b; Figure 2). Espionage and pre-positioning, however, are designed not to disrupt service and rarely appear in such statistics. Both observations point in the same direction: telecommunications security requires a resilience perspective, not only a threat-centric one.

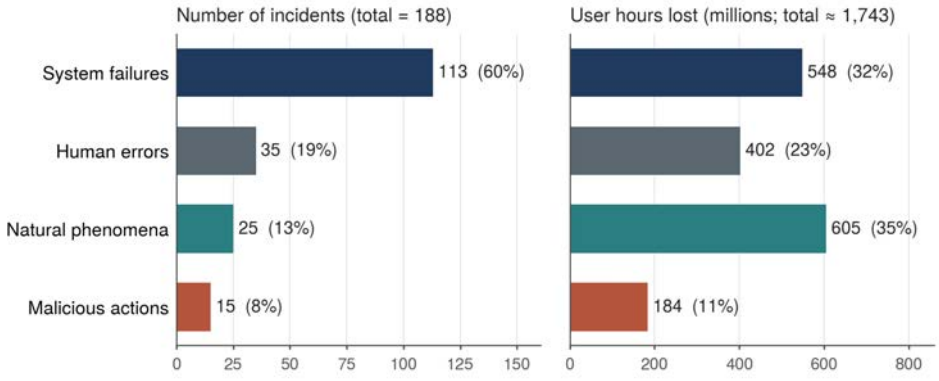

Fig. 2: Major telecommunications incidents reported in the EU/EFTA in 2024, by root cause

4. Cyber Warfare and the Telecommunications Battlefield

An adversary seeking to weaken a country may not need to attack military systems directly. Disrupting civilian telecommunications can have second-order effects on government, emergency response, finance and public confidence. The war in Ukraine provided clear evidence. On 24 February 2022, a cyberattack on the management infrastructure of Viasat's KA-SAT network disabled thousands of satellite modems, with spill-over effects across Europe (Viasat, 2022); the EU, the United States and the United Kingdom attributed it to Russia (Council of the European Union, 2022). In December 2023, a destructive attack against Kyivstar interrupted services for around 24 million subscribers attributed by Ukraine's Security Service to Russia's Sandworm group; the attackers had reportedly been inside the network since at least May 2023 (Kyiv Post, 2024).

Cyber operations against telecommunications may pursue *espionage, disruption or strategic positioning*. The third is the hardest to detect and reinforces the importance of threat hunting, behavioral analytics and privileged-access monitoring. *Cybersecurity must increasingly be designed for contested environments rather than peaceful operating assumptions.*

5. NIS2: From Compliance to Resilience

The NIS2 Directive marks a major evolution in European cybersecurity governance. Public electronic communications networks and services fall within its digital infrastructure sector, and the security and notification rules previously set by the European Electronic Communications Code have been absorbed into the new regime (Directive (EU) 2018/1972, 2018; Directive (EU) 2022/2555, 2022). For operators, NIS2 is less a new obligation than a significant raising of the bar.

Three provisions are central. Article 20 makes management bodies responsible for approving and overseeing cybersecurity risk-management measures. Article 21 sets minimum measures, from risk analysis and incident handling to business continuity, supply-chain security and cryptography. Article 23 introduces staged reporting: an early warning within 24 hours, a notification within 72 hours and a final report within one month. For essential entities, Member States must provide for maximum fines of at least €10 million or 2 per cent of worldwide annual turnover, whichever is higher.

The relevant question is therefore no longer "Are we compliant?" but "Are we resilient?". Compliance establishes a baseline; resilience is the ability to keep operating under adverse conditions. A mature NIS2 program should connect *risk management, security controls, operational resilience, crisis management and recovery* (Figure 3).

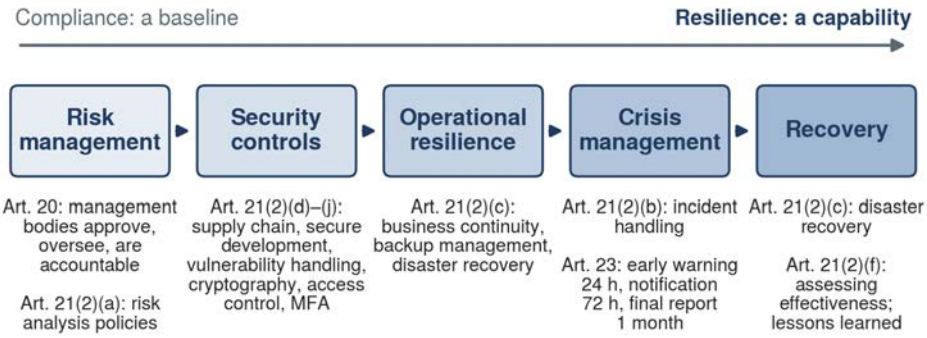

Fig. 3: From compliance to resilience: the NIS2 chain

6. 5G, 6G and the Expansion of the Trust Boundary

5G has already transformed telecommunications through virtualization, software-defined networking, network slicing and cloud-native technologies, prompting both standardized security architectures (3GPP, n.d.-a) and supplier risk assessments such as the EU 5G Toolbox (NIS Cooperation Group, 2020). 6G will go further. The ITU's IMT-2030 framework includes "AI and communication" and "integrated sensing and communication" among its usage scenarios, and lists security, privacy and resilience among the principles that apply to all of them (International Telecommunication Union – Radiocommunication Sector [ITU-R], 2023).

6G should not be understood as "faster 5G". Networks will integrate terrestrial, satellite, edge, cloud and AI components across multiple operators, becoming a federation of systems rather than a single infrastructure (Porambage et al., 2021). Who is responsible when an AI-enabled network function behaves incorrectly? Such questions must be answered at the design stage, around zero-trust principles (Rose et al., 2020). The window is narrow: 6G technology proposals are due to ITU-R in early 2029 (3GPP, n.d.-b). *Security must become an architectural property.*

7. Submarine Cables: From Security to Sovereignty

Submarine cables carry around 99 per cent of intercontinental Internet traffic (European Commission & High Representative, 2025). In the Baltic Sea, two data cables were severed within hours in November 2024 and, on 25 December 2024, the Estlink 2 power interconnector and four telecommunications cables were dam-

aged, allegedly by the dragged anchor of the tanker *Eagle S* (Reuters, 2025). A Finnish district court initially dismissed the case for lack of jurisdiction (Helsinki District Court, 2025), but the Court of Appeal overturned that ruling in September 2026, showing how contested jurisdiction over damage in international waters remains (Helsinki District Court, 2025), illustrating how hard it is to deter damage in international waters. In September 2025, cable cuts in the Red Sea degraded connectivity and cloud services between Europe, the Middle East and Asia (Al Jazeera, 2025).

Most of the 150 to 200 faults recorded worldwide each year are accidental, largely caused by fishing and anchoring (Brussels Times, 2025). The difficulty is that sabotage can be made to look accidental, slowing attribution and weakening deterrence (Bueger & Liebetrau, 2021). A cable is a physical asset, its landing stations a physical facilities, its monitoring systems digital: a classic cyber-physical problem (Figure 4).

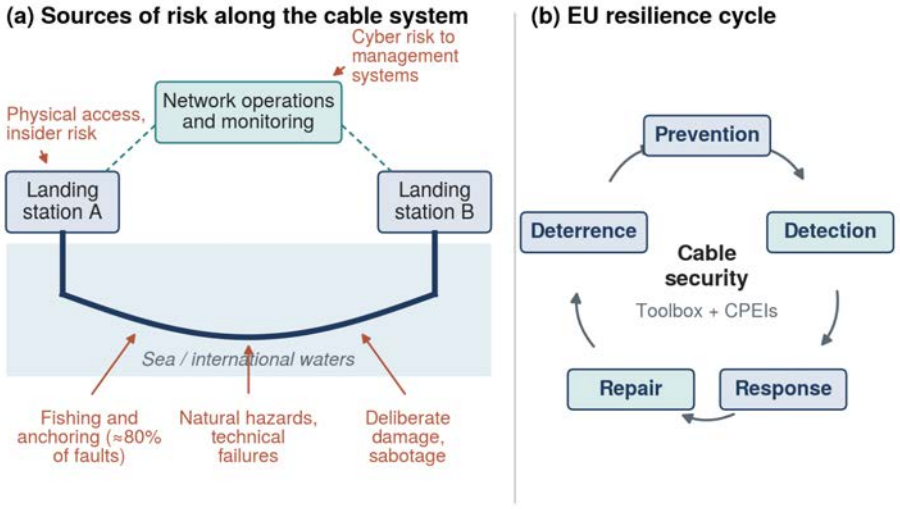

Fig. 4: Submarine cables as a cyber-physical system: sources of risk and the EU resilience cycle

The EU Action Plan on Cable Security addresses prevention, detection, response, repair and deterrence (European Commission & High Representative, 2025), and the 2026 Cable Security Toolbox adds measures on route redundancy, repair capacity and high-risk suppliers, backed by around €347 million in funding (European Commission, 2026a). The underlying analysis highlights *geographical concentration risk*: around 90 per cent of Europe–Asia data traffic transits through the Red Sea (Submarine Cable Infrastructures Expert Group, 2026). The lesson is straightforward: *redundancy is meaningful only when dependencies are genuinely independent.*

8. Satellite Connectivity

Satellites provide connectivity where terrestrial infrastructure is unavailable or damaged, which makes them essential in disasters, crises and large outages. The EU has recognized this in its Space Strategy for Security and Defence and in IRIS², its secure governmental connectivity program (European Commission & High Representative, 2023; Regulation (EU) 2023/588, 2023). Yet satellites are not an isolated backup: the KA-SAT attackers did not target the satellite but the ground network, through a misconfigured VPN appliance (Viasat, 2022). A mature strategy should combine *terrestrial* + *submarine* + *satellite* + *cloud* + *edge*, so that the failure of one layer does not cause systemic failure.

9. Quantum Computing and Crypto-Agility

Quantum computing could undermine public-key algorithms such as RSA and elliptic-curve cryptography, which protect signaling, authentication, software updates and customer communications. Estimates of when this could happen remain uncertain but are moving closer: in 2025, Gidney estimated that RSA-2048 could be factored in under a week with fewer than one million noisy qubits, twenty times fewer than his 2019 estimate (Gidney, 2025).

The risk is already present because of *harvest now, decrypt later*. Mosca (2018) expressed it as an inequality: if the time data must remain confidential plus the time needed to migrate exceeds the time until a cryptographically relevant quantum computer exists, the data are already at risk (Figure 5). Telecommunications infrastructure, with its long lifecycles, is particularly exposed.

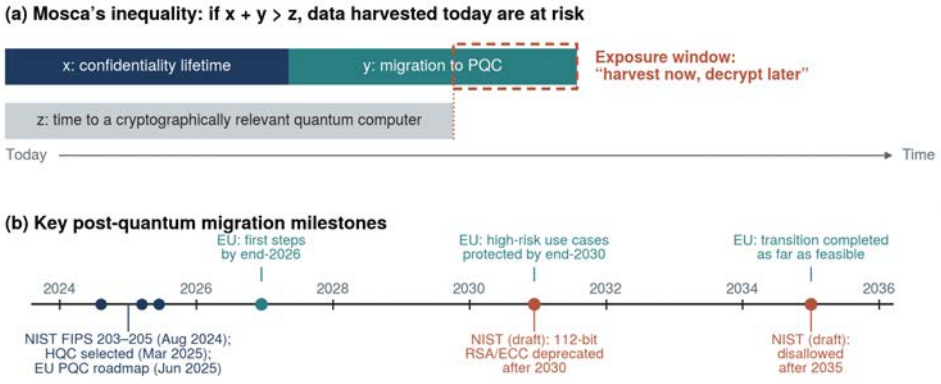

Fig. 22.5. "Harvest now, decrypt later" and the post-quantum migration timeline

NIST published its first post-quantum standards in 2024 (National Institute of Standards and Technology [NIST], 2024a, 2024b, 2024d) and has proposed to deprecate vulnerable algorithms after 2030 and disallow them after 2035 (Moody et al., 2024). The EU roadmap asks for first steps by the end of 2026 and for high-risk use cases, including critical infrastructure, to be protected by the end of 2030 (NIS Cooperation Group, 2025). For operators, the key concept is *crypto-agility*: the ability to replace algorithms without redesigning the infrastructure, starting from a complete cryptographic inventory. *Quantum resilience is not a future project. It is an infrastructure lifecycle project.*

10. Cloud Sovereignty and Managed Interdependence

Cloud infrastructure increasingly supports network functions, AI workloads and operational platforms. This brings scalability, but also strategic dependency. The question is not simply "Is the cloud secure?" but "Who controls the infrastructure, the data, the software, the keys and the legal environment?" In June 2026, the European Commission proposed the Cloud and AI Development Act, noting that EU providers' share of the European cloud market had fallen from around 29 per cent in 2017 to 15 per cent in 2022, and introducing graduated sovereignty assurance levels for public-sector procurement (European Commission, 2026c).

Sovereignty, however, should not be confused with isolation. Eliminating every foreign dependency would be unrealistic. The objective is *managed interdependence*: knowing where dependencies exist, how critical and replaceable they are, which jurisdictions apply, who controls encryption keys and what alternatives exist in a crisis. The Commission's 2026 tech sovereignty strategy frames sovereignty in similar terms (European Commission, 2026b). *The objective is not autarky. It is resilience through choice.*

11. Artificial Intelligence and Autonomous Networks

AI already supports network optimization, anomaly detection, fraud prevention and incident response. If it becomes part of the network itself, compromising the AI layer may become equivalent to compromising the network. The EU AI Act therefore classifies AI used as a safety component in critical digital infrastructure as high-risk, with the related obligations applying from December 2027 (Regulation (EU) 2024/1689, 2024; Regulation (EU) 2026/1744, 2026).

This is no longer a distant scenario. TM Forum measures network autonomy on a scale from Level 0, manual operation, to Level 5, a fully autonomous network

(TM Forum, 2023). At Level 4, most operational processes are executed by AI agents within policy constraints, with humans involved mainly in exception handling and governance (NGMN Alliance, 2026). According to TM Forum (2026), 81 per cent of operators plan to reach Level 4 or above within three to five years.

Level 4 transfers decision authority from people to software agents. A compromised agent, or manipulated intents, telemetry or models, could make the network act against itself at machine speed. Trust is not yet proven: according to a TM Forum survey, 72 per cent of operators consider their AI trustworthy, but only 14 per cent can demonstrate it (TM Forum, 2026). NGMN therefore calls for agents with verifiable identities and least privilege, guardrails, validation in digital twins and kill switches, and stresses that operators remain accountable for every action their agents take (NGMN Alliance, 2026).

In practice, Level 4 changes the role of people rather than eliminating it. Humans move from being in the loop, approving each action, to being on the loop: they define intents and policies, supervise outcomes and intervene by exception (NGMN Alliance, 2026). This makes the supervisory layer itself critical. One emerging approach is the use of custodian/guardian/keeper agents, AI-based components that review and monitor the actions of other agents and can redirect or block those that do not align with predefined goals and policies. Oversight, in other words, must itself be engineered, secured and tested. *Autonomy without governance is not resilience.*

12. Supply Chain Security

An operator depends on thousands of hardware, software and service components, and a compromised supplier can open access to otherwise well-protected systems. The faulty CrowdStrike update of July 2024, which crashed around 8.5 million Windows devices (Weston, 2024), and the backdoor discovered in the XZ Utils library (CISA, 2024) illustrate the accidental and the deliberate versions of the same risk. *The security perimeter extends to the supply chain.* European policy reflects this through the Cyber Resilience Act (Regulation (EU) 2024/2847, 2024) and the proposed Cybersecurity Act 2, which would restrict high-risk suppliers of key ICT assets, with a phase-out period for mobile networks (European Commission, 2026d).

13. From Prevention to Continuity

Prevention remains essential but is not sufficient. A resilient operator must assume that sophisticated attacks may eventually succeed and ask: “Can the

organization continue operating when prevention fails?" NIST defines cyber resiliency as the ability to anticipate, withstand, recover from and adapt to adverse conditions (Ross et al., 2021). This requires mapping critical services and dependencies, testing recovery, keeping alternative communication channels and, as the Optus review showed, adequate out-of-band management (Bean, 2024). A five-minute outage and a five-hour outage are fundamentally different events. Figure 6 illustrates this with the resilience curve, a representation originally developed for infrastructure exposed to natural disasters (Bruneau et al., 2003). The objective is to limit both the depth of the performance loss and the time needed to recover, which together determine the overall impact, and to emerge from each disruption better adapted than before.

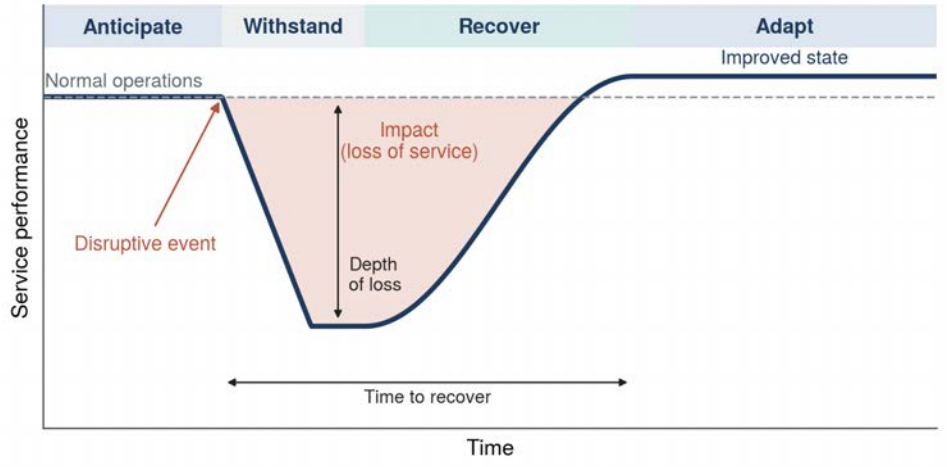

Fig. 22.6: The resilience curve: anticipate, withstand, recover, adapt

A major incident is rarely only technical: it is also operational, regulatory, reputational and potentially a national-security event. All functions, up to executive management, must act as a single crisis-management system, connected to national and EU arrangements such as the 2025 EU blueprint for cyber crisis management (Council of the European Union, 2025). Exercises should combine scenarios such as ransomware with network disruption, the loss of a critical supplier, a cable cut or a misbehaving autonomous agent. *The objective is not simply to test technology. It is to test leadership.*

Resilience also requires public-private cooperation, since each actor holds part of the threat picture, and it must be measured: time to detect, respond and recover; how long critical services can run in degraded mode; the real independence of alternative routes; cryptographic readiness; and whether governance keeps pace with automation.

14. A Multi-Layer Resilience Framework and the Role of the CISO

These elements can be combined into seven interconnected layers: network security; cloud and edge security; physical connectivity; space connectivity; cryptographic security; supply-chain security; and strategic and national resilience (Figure 7). They cannot be managed independently. A resilient network requires resilience across all seven.

This changes the role of the CISO, who becomes a strategic risk leader able to translate technical risk into strategic consequences, consistent with NIS2's emphasis on management accountability and with the Govern function of the NIST Cybersecurity Framework 2.0 (NIST, 2024c). A cable vulnerability is a connectivity risk; a cloud dependency is a sovereignty risk; a quantum-vulnerable algorithm is a long-term confidentiality risk; a 6G architecture may redefine the organization's entire trust model.

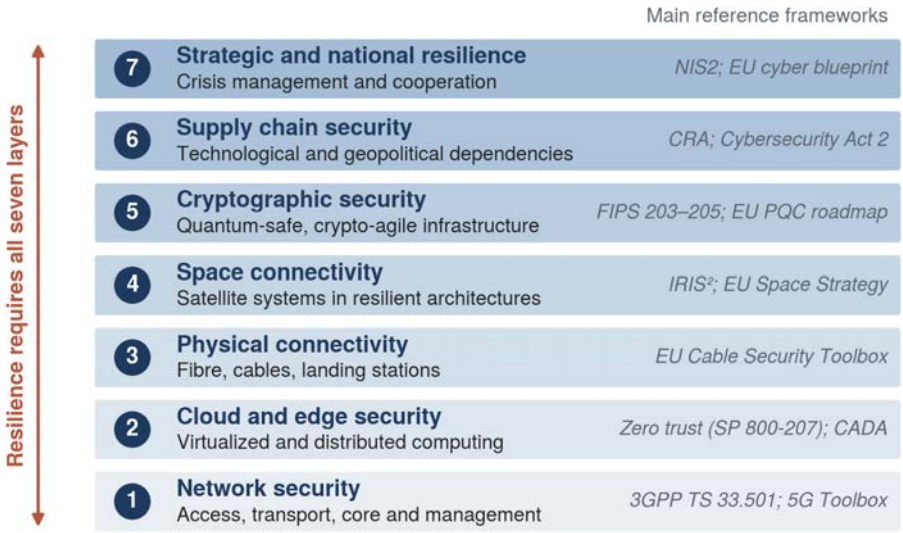

Fig. 22.7: A multi-layer telecommunications resilience framework

15. Strategic Recommendations

Operators and policymakers should consider eight priorities: (1) treat telecommunications as strategic infrastructure; (2) use NIS2 as a catalyst for resilience, not only for compliance; (3) design 6G and autonomous operations for security from the outset; (4) bring cables, landing stations and other physical

assets into cyber-resilience strategies; (5) build complementary terrestrial, submarine, satellite, cloud and edge layers; (6) begin quantum migration with cryptographic inventories and crypto-agility; (7) assess suppliers and cloud providers for geopolitical and concentration risk as well as cybersecurity risk; and (8) strengthen public-private information sharing and coordinated crisis response.

16. Conclusion: Securing the Backbone of Society

Telecommunications are no longer simply the infrastructure that connects people. They are the infrastructure that enables society to function. 5G and 6G, cloud and edge, AI, satellites and submarine cables are turning them into a complex multi-layer critical infrastructure, while quantum computing challenges their cryptographic foundations and cyber warfare blurs the boundary between civilian and national security.

The CISO of the future will need to understand not only cyber threats, but also geopolitical risk, technological sovereignty, supply-chain dependencies, space infrastructure, quantum computing and national resilience.

The question is not whether an attacker will eventually succeed, but whether that compromise can become systemic. NIS2 provides an important foundation, but regulation alone cannot create resilience. Resilience is created through architecture, investment, governance, people and leadership.

Europe's objective should be a telecommunications ecosystem that is innovative without being fragile, interconnected without being excessively dependent, and open without becoming strategically vulnerable. Technological sovereignty should not mean isolation. Redundancy should not mean duplication without independence. Compliance should not become bureaucracy.

The backbone of the digital society must therefore be more than connected. *It must be secure, resilient, sovereign where strategically necessary, and capable of sustaining society through disruption.*

References

3GPP. (n.d.-a). *Security architecture and procedures for 5G system* (3GPP TS 33.501). 3rd Generation Partnership Project. https://www.3gpp.org/dynareport/33501.htm

3GPP. (n.d.-b). *Release 20*. 3rd Generation Partnership Project. Retrieved September 30, 2026, from https://www.3gpp.org/specifications-technologies/releases/release-20

Al Jazeera. (2025, September 7). *Internet disruptions in Middle East and South Asia after Red Sea cable cuts*. https://www.aljazeera.com/news/2025/9/7/internet-disruptions-in-middle-east-and-south-asia-after-red-sea-cable-cuts

Bean, R. (2024). *Review into the Optus outage of 8 November 2023: Final report*. Australian Government, Department of Infrastructure, Transport, Regional Development, Communications and the Arts. https://www.infrastructure.gov.au/department/media/publications/review-optus-outage-8-november-2023-final-report

Brussels Times. (2025, February 18). *Fishing more likely to be behind damage to underwater cables than spies, says UN*. https://www.brusselstimes.com/world/1448914/fishing-not-spies-more-likely-to-be-behind-damage-to-underwater-cables-says-un

Bruneau, M., Chang, S. E., Eguchi, R. T., Lee, G. C., O'Rourke, T. D., Reinhorn, A. M., Shinozuka, M., Tierney, K., Wallace, W. A., & von Winterfeldt, D. (2003). *A framework to quantitatively assess and enhance the seismic resilience of communities*. Earthquake Spectra, 19(4), 733–752.

Bueger, C., & Liebetrau, T. (2021). Protecting hidden infrastructure: The security politics of the global submarine data cable network. *Contemporary Security Policy, 42*(3), 391–413. https://doi.org/10.1080/13523260.2021.1907129

Council of the European Union. (2022, May 10). *Russian cyber operations against Ukraine: Declaration by the High Representative on behalf of the European Union* https://www.consilium.europa.eu/en/press/press-releases/2022/05/10/russian-cyber-operations-against-ukraine-declaration-by-the-high-representative-on-behalf-of-the-european-union/.

Council of the European Union. (2025). Council Recommendation of 6 June 2025 on an EU blueprint for cyber crisis management (C/2025/3445). *Official Journal of the European Union*, C. http://data.europa.eu/eli/C/2025/3445/oj

Cybersecurity and Infrastructure Security Agency. (2024, March 29). *Reported supply chain compromise affecting XZ Utils data compression library, CVE-2024-3094* [Alert]. https://www.cisa.gov

Cybersecurity and Infrastructure Security Agency, National Security Agency, & Federal Bureau of Investigation. (2024). *PRC state-sponsored actors compromise and maintain persistent access to U.S. critical infrastructure* (Cybersecurity Advisory AA24-038A). https://www.cisa.gov/news-events/cybersecurity-advisories/aa24-038a

Cybersecurity and Infrastructure Security Agency, National Security Agency, Federal Bureau of Investigation, et al. (2025). *Countering Chinese state-sponsored actors compromise of networks worldwide to feed global espionage system* (Cybersecurity Advisory AA25-239A). https://www.cisa.gov/news-events/alerts/2025/08/27/cisa-and-partners-release-joint-advisory-countering-chinese-state-sponsored-actors-compromise

Directive (EU) 2018/1972 of the European Parliament and of the Council of 11 December 2018 establishing the European Electronic Communications Code. (2018). *Official Journal of the European Union*, L 321, 36–214.

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 (NIS 2 Directive). (2022). *Official Journal of the European Union*, L 333, 80–152.

European Commission. (2026a, February 11). *Commission increases submarine cable security with €347 million investment* [Press release].

European Commission. (2026b). *Communication on European tech sovereignty, accompanied by an EU open source strategy* (COM(2026) 503 final).

European Commission. (2026c). *Proposal for a regulation of the European Parliament and of the Council establishing a framework of measures for strengthening Europe's cloud and AI ecosystem (Cloud and AI Development Act)* (COM(2026) 502 final).

European Commission. (2026d). *Proposal for a regulation of the European Parliament and of the Council on the European Union Agency for Cybersecurity (ENISA), European cybersecurity certification framework, and ICT supply chain security and repealing Regulation (EU) 2019/881 (The Cybersecurity Act 2)* (COM(2026) 11 final).

European Commission & High Representative of the Union for Foreign Affairs and Security Policy. (2023). *European Union space strategy for security and defence* (JOIN(2023) 9 final).

European Commission & High Representative of the Union for Foreign Affairs and Security Policy. (2025). *EU action plan on cable security* (JOIN(2025) 9 final).

European Union Agency for Cybersecurity. (2025a). *ENISA threat landscape 2025*. ENISA. https://www.enisa.europa.eu

European Union Agency for Cybersecurity. (2025b). *Telecom security incidents 2024*. ENISA. https://www.enisa.europa.eu/publications/telecom-security-incidents-2024

Giannopoulos, G., Smith, H., & Theocharidou, M. (2021). *The landscape of hybrid threats: A conceptual model* (Public version). Publications Office of the European Union.

Gidney, C. (2025). *How to factor 2048 bit RSA integers with less than a million noisy qubits* (arXiv:2505.15917). arXiv. https://arxiv.org/abs/2505.15917

Helsinki District Court. (2025, October 3). *Judgment in the Eagle S case* [English translation].

International Telecommunication Union – Radiocommunication Sector. (2023). *Framework and overall objectives of the future development of IMT for 2030 and beyond* (Recommendation ITU-R M.2160-0). ITU.

Kyiv Post. (2024, January 4). *Russian hackers infiltrated Kyivstar mobile provider system six months prior to cyberattack*. https://www.kyivpost.com/post/26313

Le Maistre, R. (2025, May 1). *Iberian outage highlights power security challenges for network operators*. TelecomTV. https://www.telecomtv.com/content/security/iberian-outage-highlights-power-security-challenges-for-network-operators-52926/

Moody, D., Perlner, R., Regenscheid, A., Robinson, A., & Cooper, D. (2024). *Transition to post-quantum cryptography standards* (NIST IR 8547, initial public draft). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8547.ipd

Mosca, M. (2018). Cybersecurity in an era with quantum computers: Will we be ready? *IEEE Security & Privacy, 16*(5), 38–41. https://doi.org/10.1109/MSP.2018.3761723

National Institute of Standards and Technology. (2024a). *Module-lattice-based digital signature standard* (FIPS 204). https://doi.org/10.6028/NIST.FIPS.204

National Institute of Standards and Technology. (2024b). *Module-lattice-based key-encapsulation mechanism standard* (FIPS 203). https://doi.org/10.6028/NIST.FIPS.203

National Institute of Standards and Technology. (2024c). *The NIST Cybersecurity Framework (CSF) 2.0* (NIST CSWP 29). https://doi.org/10.6028/NIST.CSWP.29

National Institute of Standards and Technology. (2024d). *Stateless hash-based digital signature standard* (FIPS 205). https://doi.org/10.6028/NIST.FIPS.205

National Institute of Standards and Technology. (2025, March 11). *NIST selects HQC as fifth algorithm for post-quantum encryption* [News release].

NGMN Alliance. (2026). *Network automation and autonomy phase III: Agentic AI for autonomous mobile networks* (Version 1.0). https://www.ngmn.org/wp-content/uploads/2608_NGMN_Network-Automation-and-Autonomy-Phase-III_v1.0.pdf

NIS Cooperation Group. (2020). *Cybersecurity of 5G networks: EU toolbox of risk mitigating measures* (CG Publication 01/2020).

NIS Cooperation Group. (2025). *A coordinated implementation roadmap for the transition to post-quantum cryptography*. https://digital-strategy.ec.europa.eu/en/library/coordinated-implementation-roadmap-transition-post-quantum-cryptography

Porambage, P., Gür, G., Moya Osorio, D. P., Liyanage, M., Gurtov, A., & Ylianttila, M. (2021). The roadmap to 6G security and privacy. *IEEE Open Journal of the Communications Society, 2*, 1094–1122.

Regulation (EU) 2023/588 of the European Parliament and of the Council of 15 March 2023 establishing the Union Secure Connectivity Programme for the period 2023–2027. (2023). *Official Journal of the European Union*, L 79, 1–39.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). *Official Journal of the European Union*, L, 2024/1689.

Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act). (2024). *Official Journal of the European Union*, L, 2024/2847.

Regulation (EU) 2026/1744 of the European Parliament and of the Council of 8 July 2026 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI). (2026). *Official Journal of the European Union*, L, 2026/1744.

Reuters. (2025, February 21). A timeline of suspected underwater sabotage incidents in the Baltic Sea. *gCaptain*. https://gcaptain.com/a-timeline-of-suspected-underwater-sabotage-incidents-in-the-baltic-sea/

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). *Zero trust architecture* (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207

Ross, R., Pillitteri, V., Graubart, R., Bodeau, D., & McQuaid, R. (2021). *Developing cyber-resilient systems: A systems security engineering approach* (NIST Special Pub-

lication 800-160 Vol. 2 Rev. 1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-160v2r1

Submarine Cable Infrastructures Expert Group. (2026). *Report on cable security toolbox of mitigation measures and cable projects of European interest*. European Commission, Directorate-General for Communications Networks, Content and Technology.

TM Forum. (2023). *Autonomous network levels evaluation methodology* (IG1252, Version 1.2.0). https://www.tmforum.org/resources/standard/ig1252-autonomous-network-levels-evaluation-methodology-v1-2-0

TM Forum. (2026). *TM Forum launches Race to 2030 at DTW Ignite 2026* [Press release]. https://www.tmforum.org/news-insight/newsroom/tm-forum-launches-race-to-2030-at-dtw-ignite-2026

Viasat. (2022, March 30). *KA-SAT network cyber attack overview*. https://news.viasat.com/blog/corporate/ka-sat-network-cyber-attack-overview

Weston, D. (2024, July 20). *Helping our customers through the CrowdStrike outage*. Official Microsoft Blog. https://blogs.microsoft.com/blog/2024/07/20/helping-our-customers-through-the-crowdstrike-outage/

SPACE SYSTEMS AS CRITICAL INFRASTRUCTURES: CYBERSECURITY CHALLENGES FOR EUROPEAN RESILIENCE AND STRATEGIC SOVEREIGNTY

Daniele Frasca
Thales Alenia Space Italia, Italy

Vittoria Piantelli
Thales Alenia Space Italia, Italy

Abstract
Modern space systems have evolved from basic technological platforms into critical strategic infrastructures essential for governmental services, commercial and defense, and societal functions in Europe. The increasing reliance on space-based services amid complex geopolitical challenges highlights their importance for European resilience and sovereignty. They operate as integrated mission ecosystems combining Space and Ground segments, Secure Communications, Operations, and User Services, designed for autonomous, long-term operation in hostile environments. Ensuring resilience from the early design stages is vital, emphasizing reliability, safety, security, and product assurance throughout the entire lifecycle, as space systems cannot rely on post-deployment in-orbit repairs. Cybersecurity is embedded throughout the system lifecycle to protect mission continuity, integrity and trustworthiness. Approaches like Security by Design, Mission Assurance, and recognized frameworks such as Common Criteria and European regulations like NIS2 and the Cyber Resilience Act, support the system resilience. Protecting space systems is a complex engineering challenge, given their role as critical infrastructures and emergency management, vital for safeguarding Europe's resilience, continuity and strategic autonomy.

1. Introduction – Space as a Critical Domain

Space has become an essential domain for the functioning of contemporary societies. Satellite communications, navigation and timing, Earth observation and other space-based capabilities increasingly support economic activities, public services, critical infrastructures, security and defense. What was once a highly specialized technological domain is therefore becoming an integral part of the infrastructure on which modern societies depend.

This growing reliance on space also creates new forms of vulnerability. Space systems are complex, interconnected and increasingly integrated with terrestrial digital infrastructures. A disruption affecting a space-based capability may consequently generate effects extending far beyond the space system itself, potentially affecting multiple services and sectors simultaneously. Cybersecurity must

therefore be considered not only in relation to the protection of individual space assets, but also in terms of the broader resilience of the functions they enable.

For Europe, this transformation has technological, regulatory, industrial and strategic implications. This contribution argues that space systems should be understood not only as critical infrastructures, but also as critical enablers of other critical infrastructures. From this perspective, it examines the distinctive cybersecurity challenges of space systems, the evolution from regulatory compliance towards cyber resilience, and the relationship between dual-use capabilities, industrial competence and European strategic sovereignty. Ultimately, securing space means protecting an increasingly fundamental component of Europe's resilience and capacity for autonomous action.

2. Space Systems as Critical Infrastructures – and Critical Enablers

Space systems have evolved from specialized technological assets into essential components of modern societies. Satellite communications, navigation and positioning, timing services and Earth observation support an increasingly broad range of economic, institutional and security-related activities. Their relevance can therefore no longer be assessed exclusively in terms of the value of the satellites themselves or of the services directly delivered to space-sector users.

A space system is, moreover, much more than a spacecraft. It is a complex and distributed infrastructure comprising the space segment, ground control and mission segments, communication links, command-and-control functions, data-processing facilities and interfaces with users and service providers. Security consequently concerns this entire interconnected architecture. This end-to-end perspective is particularly relevant because the compromise or unavailability of one element may affect the operation of the overall system. Industrial experience with complex space architectures has long reflected this system-level approach, integrating protection of data, assets and operations throughout the different segments of a mission.

However, considering space systems simply as critical infrastructures captures only part of their contemporary strategic relevance. Space-based capabilities increasingly act as critical enablers of other critical infrastructures and essential services. Navigation and timing support transportation, telecommunications, energy and financial systems; satellite communications enable governmental, commercial and emergency connectivity; Earth observation contributes to environmental monitoring, disaster management, security and defense. The resulting dependency therefore extends well beyond the space domain itself.

This interdependence introduces the possibility of cascading and systemic effects. A disruption affecting the availability, integrity or authenticity of a space-based service may propagate to terrestrial infrastructures whose operation de-

pends upon that service. Protecting space systems therefore means not only protecting the space infrastructure itself, but also contributing to the protection and continuity of the infrastructures and essential functions that depend on space.

This distinction has important implications for cybersecurity. The security of space systems becomes part of the resilience of a broader ecosystem of interconnected infrastructures, rather than a sector-specific technological concern. This perspective is particularly relevant for Europe, where space capabilities increasingly support civilian, commercial, governmental and security-related functions. Understanding space systems simultaneously as critical infrastructures and critical enablers thus provides the conceptual basis for addressing cybersecurity as an essential component of European resilience.

3. The Specific Cybersecurity Challenges of Space Systems

The cybersecurity of space systems cannot simply replicate approaches developed for conventional terrestrial information systems. Space missions combine technological, operational and lifecycle constraints that create a distinctive security environment. Their protection therefore requires an end-to-end perspective capable of addressing not only individual components, but also the interactions among space, ground and communication infrastructures.

A first challenge derives from the distributed nature of space architectures. Satellites operate as part of complex systems involving ground stations, mission and control centers, communication links, data-processing facilities and user segments. The attack surface consequently extends across multiple interconnected environments and includes both conventional cyber threats and threats specific to space communications. Interception, spoofing, replay and intentional interference such as jamming may affect communication links, while vulnerabilities in ground infrastructures or interfaces may potentially compromise the wider mission. Security cannot therefore be confined to the protection of the spacecraft itself.

A second distinctive characteristic is the limited possibility of physical intervention on an asset once deployed in orbit. Terrestrial infrastructures can normally be repaired, upgraded or replaced; comparable interventions on a spacecraft may be extremely difficult or impossible. Hardware architectures and security mechanisms must consequently remain reliable throughout missions lasting many years. This creates a further challenge: the threat environment evolves considerably faster than the systems designed to withstand it. Technologies and cryptographic mechanisms considered sufficiently robust during development may face more capable attackers and new attack techniques during the operational lifetime of the mission. Cybersecurity decisions must therefore consider not only current threats, but also the long-term resilience of the system.

These constraints also modify the traditional balance between confidentiality, integrity and availability. Confidentiality remains essential for sensitive information and integrity is fundamental for command, control and mission data. However, availability frequently becomes a particularly significant driver in space systems. Loss of communications, denial of command capabilities or disruption of critical functions may compromise the mission and, as discussed above, affect terrestrial services dependent upon space-based capabilities. Security engineering must therefore combine protection against unauthorized access and manipulation with mechanisms supporting continuity, fault tolerance and secure recovery.

The long development and operational lifecycle of space systems has another important consequence: cybersecurity cannot effectively be added after the architecture has been defined. Threat identification, risk analysis and security requirements need to influence system design, implementation, verification and validation from the earliest phases. In this context, security becomes an engineering process extending throughout the mission lifecycle rather than a collection of individual technological countermeasures. Industrial space-security practice reflects this principle through the concept that threats should be "countered by design", with the effectiveness of security measures subsequently demonstrated through appropriate testing, verification and assurance activities.

Finally, the increasing convergence of civilian, commercial, governmental and defense applications, together with complex international supply chains, further expands the cybersecurity challenge. Space security consequently involves technology, architecture, operations, organizations and trust across the entire industrial ecosystem.

These characteristics point towards a broader concept of cyber resilience. Protecting a space system means not only preventing attacks, but ensuring that essential functions can remain trustworthy and available under evolving threats. The question for Europe is therefore how regulatory requirements, standards and assurance mechanisms can support this lifecycle-oriented approach without reducing cybersecurity to formal compliance.

4. From Compliance to Cyber Resilience: The European Regulatory Perspective

The European cybersecurity framework increasingly recognizes that resilience cannot be achieved through isolated technical measures. For the space sector, this evolution is particularly significant because regulatory obligations must address infrastructures characterized by long lifecycles, complex supply chains and strong interdependencies with other critical sectors.

The NIS2 Directive represents an important step by explicitly recognizing

"Space as a sector of high criticality". However, its space-specific scope is focused on operators of ground-based infrastructure supporting the provision of space-based services, rather than on the space system as a whole. The Italian transposition through Legislative Decree No. 138/2024 reflects the same approach. This regulatory perimeter is particularly significant from a cybersecurity perspective: while it acknowledges the criticality of the terrestrial infrastructure enabling space services, the engineering challenge remains inherently end-to-end, encompassing space, ground and communication segments throughout the mission lifecycle.

A complementary perspective is provided by the Cyber Resilience Act (CRA), which establishes horizontal cybersecurity requirements for products with digital elements. The Regulation emphasizes cybersecurity throughout product design, development and production, together with vulnerability handling during the expected period of use. This lifecycle approach is highly relevant to the space sector, although its applicability must be assessed according to the characteristics, intended purpose and regulatory status of the specific product, with dedicated exclusions notably applying to products developed exclusively for national security or defense purposes.

The direction of European policy becomes even clearer in the Commission's 2025 proposal for an EU Space Act. Its resilience pillar introduces space-specific cybersecurity requirements intended to complement the Union's horizontal cybersecurity framework. Significantly, the proposal addresses risk management throughout the lifecycle of space missions and considers system and component risks, cryptography and key management, incident response, business continuity and security testing.

These regulatory developments converge with principles that have long characterized security engineering for high-assurance space systems. Standards and assurance frameworks, including NIST guidance, CCSDS security standards and the Common Criteria, provide methodologies for translating security objectives into technical requirements and for demonstrating, through structured evaluation and verification, that defined security claims have been adequately addressed. Industrial practice similarly shows that security effectiveness depends on integrating risk analysis, design, implementation, verification and assurance throughout the system lifecycle.

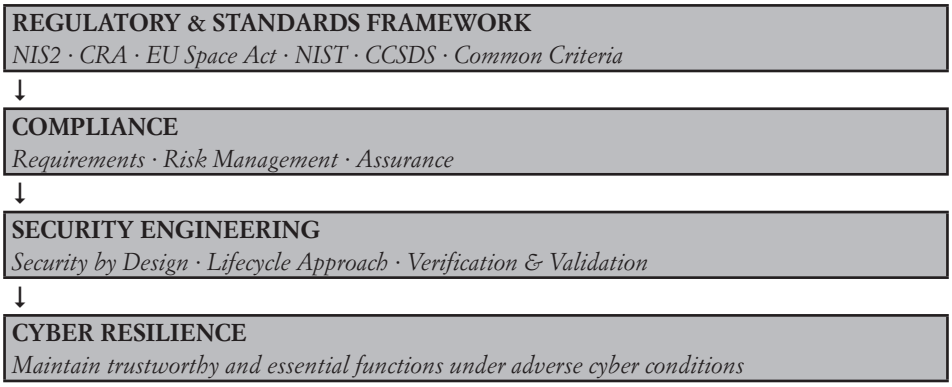

Fig. 1: From regulatory compliance to cyber resilience in space systems

However, compliance should be regarded as a baseline rather than the final objective. A system may formally satisfy applicable requirements while remaining insufficiently prepared for evolving threats, unexpected failure modes or complex dependencies. Cyber resilience requires moving beyond the question "Are the requirements satisfied?" towards a more demanding one: "Can the system continue to deliver its essential functions under adverse cyber conditions?"

For European space systems, regulation, standards, certification and engineering assurance should therefore operate as complementary instruments within a common lifecycle-oriented approach, transforming cybersecurity from a formal obligation into a measurable property of trustworthy and resilient space systems.

5. European Resilience, Dual Use and Strategic Sovereignty

The increasing dependence of European societies on space-based services gives cybersecurity a strategic dimension that extends beyond the protection of individual systems. If space capabilities support critical infrastructures, governmental functions and essential services, their resilience becomes directly connected to Europe's capacity to maintain technological, operational and industrial capabilities under its own control.

This relationship is particularly evident in the dual-use nature of space technologies. Satellite communications, navigation, Earth observation and secure data services may simultaneously support civilian, commercial, governmental and defense applications. The distinction between these domains is therefore increasingly less relevant from the perspective of technological dependency: the same enabling technologies, industrial capabilities and infrastructures may con-

tribute to services with very different missions and security requirements. Industrial space-security experience has long reflected this characteristic through architectures designed to support multiple user communities while maintaining appropriate separation, protection and assurance mechanisms.

Dual use also changes the meaning of cyber resilience. A vulnerability affecting a technology, component or service may have consequences across several operational domains. Conversely, the ability to design and maintain trusted technologies can provide resilience simultaneously to civilian and security-related applications. Cybersecurity therefore becomes not only a protective function, but an enabling capability for the trusted use of space technologies across different strategic domains.

This leads directly to the question of European strategic sovereignty. Sovereignty in this context should not be interpreted as technological isolation or complete autonomy from global supply chains. Modern space systems are intrinsically international and depend on highly specialized industrial ecosystems. Strategic sovereignty should rather be understood as Europe's ability to retain sufficient control over the critical technologies, competencies and industrial capabilities required to understand risks, make independent security decisions and ensure the continuity of essential space-based services.

From this perspective, cybersecurity sovereignty depends not only on possessing technologies, but also on retaining the capability to design, evaluate, certify and operate them with an adequate level of assurance. Cryptographic technologies, key-management capabilities and security evaluation provide particularly clear examples: trust cannot rely exclusively on the availability of a product, but also requires knowledge of its security properties and the ability to assess whether those properties satisfy mission requirements. European industrial capabilities in high-assurance security engineering and certification therefore represent strategic assets in themselves.

The objective should consequently not be technological self-sufficiency, but the reduction of strategic dependencies that could limit Europe's freedom of action or undermine the resilience of critical services. European initiatives aimed at strengthening secure connectivity, autonomous space capabilities and a competitive industrial base should be considered within this broader perspective.

Cyber resilience, industrial capability and strategic sovereignty are therefore mutually reinforcing. A resilient European space ecosystem requires not only regulation and secure technologies, but also the knowledge, skills and industrial capacity necessary to control their critical security properties throughout the lifecycle. Cybersecurity thus becomes an integral component of European strategic sovereignty rather than merely a technical requirement of individual space programs.

6. Conclusions – Towards a European Space Cybersecurity Culture

Space systems have become essential components of the infrastructure on which European societies, economies and institutions increasingly depend. Their strategic relevance derives not only from their intrinsic criticality, but also from their role as critical enablers of other critical infrastructures and essential services. Consequently, the cybersecurity of space systems can no longer be considered a sector-specific technical issue.

The distinctive characteristics of the space environment reinforce this conclusion. Long mission lifecycles, limited possibilities for physical intervention, distributed architectures and evolving threats require cybersecurity to be integrated into the entire system lifecycle. Regulation, standards and certification provide indispensable foundations, but compliance alone does not create resilience. Security must be designed, verified and continuously managed as an engineering property of the system.

For Europe, this challenge also has a strategic dimension. The ability to design, evaluate, protect and operate trusted space technologies contributes directly to resilience and reduces dependencies that could constrain European freedom of action. Cybersecurity, industrial capability and strategic sovereignty are therefore increasingly interconnected.

Europe should consequently move towards a genuine space cybersecurity culture, bringing together regulation, security engineering, industrial competence and strategic awareness. The ultimate objective is not merely to protect individual space assets, but to ensure that Europe can continue to rely on trustworthy space-based services under evolving cyber conditions. In this perspective, cyber resilience becomes both an engineering requirement and a strategic capability.

References

European Parliament and Council (2022). Directive (EU) 2022/2555 of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS2 Directive). Official Journal of the European Union, L 333, 27 December 2022, pp. 80–152.
https://eur-lex.europa.eu/eli/dir/2022/2555/oj
Italian Republic (2024). Decreto legislativo 4 settembre 2024, n. 138. Recepimento della direttiva (UE) 2022/2555. Gazzetta Ufficiale della Repubblica Italiana, Serie Generale n. 230, 1 October 2024.
https://www.gazzettaufficiale.it/eli/id/2024/10/01/24G00155/SG
European Parliament and Council (2024). Regulation (EU) 2024/2847 of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements (Cyber Resilience Act). Official Journal of the European Union, 20 November 2024.
https://eur-lex.europa.eu/eli/reg/2024/2847/oj
European Commission and High Representative of the Union for Foreign Affairs and Security Policy (2023). European Union Space Strategy for Security and Defence. Joint Communication, JOIN(2023) 9 final, 10 March 2023.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52023JC0009
European Commission (2025). Proposal for a Regulation of the European Parliament and of the Council on the safety, resilience and sustainability of space activities in the Union (EU Space Act). COM(2025) 335 final, 25 June 2025. Legislative proposal.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52025PC0335
National Institute of Standards and Technology (2022). Engineering Trustworthy Secure Systems. NIST Special Publication 800-160, Volume 1, Revision 1.
https://doi.org/10.6028/NIST.SP.800-160v1r1
National Institute of Standards and Technology (2021). Developing Cyber-Resilient Systems: A Systems Security Engineering Approach. NIST Special Publication 800-160, Volume 2, Revision 1.
https://doi.org/10.6028/NIST.SP.800-160v2r1
National Institute of Standards and Technology (2024). The NIST Cybersecurity Framework (CSF) 2.0. NIST Cybersecurity White Paper 29.
https://doi.org/10.6028/NIST.CSWP.29
Consultative Committee for Space Data Systems (2012). Security Architecture for Space Data Systems. CCSDS 351.0-M-1, Recommended Practice, Issue 1, November 2012.
https://ccsds.org/Pubs/351x0m1.pdf
Consultative Committee for Space Data Systems (2022). Space Data Link Security Protocol. CCSDS 355.0-B-2, Recommended Standard, Issue 2.
https://ccsds.org/Pubs/355x0b2.pdf
Common Criteria (2022). Common Criteria for Information Technology Security Evaluation. CC:2022, Revision 1, Parts 1–5, November 2022.
https://www.commoncriteriaportal.org/cc/
European Cooperation for Space Standardization (2017). Space engineering – System engineering general requirements. ECSS-E-ST-10C Rev.1, 15 February 2017.

https://ecss.nl/standard/ecss-e-st-10c-rev-1-system-engineering-general-requirements-15-february-2017/
European Cooperation for Space Standardization (2016). Space product assurance – Product assurance management. ECSS-Q-ST-10C Rev.1, 15 March 2016.
https://ecss.nl/standard/ecss-q-st-10c-rev-1-product-assurance-management-15-march-2016/

Finito di stampare nel mese di novembre 2026
dalla *Grafica Elettronica* - Napoli

Cybersecurity and artificial intelligence are reshaping the exercise of public authority, the continuity of essential services, and the protection of individual rights. Understanding these developments requires a holistic perspective that connects law, governance, strategy, technology, and operational practice.

Cybersecurity & Artificial Intelligence: The Playbook on European Governance, Strategies, and Law places this integrated approach at its core.

The volume follows a progression from institutional authority and regulatory frameworks to international cooperation, emerging threats, and resilience. Its chapters examine cybersecurity strategies, cyber diplomacy, responsible State behaviour, cybercrime investigations, digital evidence, threat intelligence, and AI governance, including autonomous agents. Analyses of industry, healthcare, maritime operations, telecommunications, and space demonstrate how these interconnected issues arise in practice.

Bringing together expertise from academia, international institutions, law enforcement, diplomacy, and industry, the Playbook provides scholars, students, policymakers, legal practitioners, and cybersecurity professionals with a shared foundation for informed decisions and cooperation.

Developed by the University of Bari Aldo Moro in collaboration with the Luxembourg House of Cybersecurity, this open-access volume is funded by the European Union within the framework of the Jean Monnet Module "Cybersecurity in the Digital Era: European Perspectives, CdeEP."

ISBN 979-12-235-0849-7

Euro 40,00