Article | Received 4 April 2026; Revised 30 May 2026; Accepted 15 June 2026; Published X July 2026
https://doi.org/10.55092/let20260009

The regulatory trilemma of AI-driven cybersecurity in the European Union: reconciling the AI Act, DORA, and fundamental rights

Fabian Teichmann[1,*] and Bruno Sergi[2]

[1] LSE IDEAS, The London School of Economics and Political Science, London, UK
[2] Davis Center for Russian and Eurasian Studies, Harvard University, Cambridge, USA

* Correspondence author; E-mail: f.m.teichmann@lse.ac.uk.

Highlights:

- Maps a regulatory trilemma between the EU AI Act, DORA, and EU fundamental rights.
- Shows why the three tensions cannot be resolved pairwise for high-risk systems.
- Proposes calibrated autonomy: threat severity, data sensitivity, reversibility.
- Tiered protocol reserves human oversight for irreversible, consequential actions.
- Reform agenda: ESA guidance, ENISA taxonomy, narrow safe harbour, advisory boards.

Abstract: The European Union has enacted two landmark frameworks that impose partially divergent obligations on financial entities deploying artificial intelligence (AI) in cybersecurity. The AI Act (Regulation (EU) 2024/1689) establishes a risk-based classification system subjecting AI systems to graduated transparency, explainability, and human-oversight duties. The Digital Operational Resilience Act (DORA, Regulation (EU) 2022/2554) requires financial entities to maintain robust Information and Communication Technology (ICT) risk-management capabilities, including rapid, automation-capable threat detection and incident response. This article argues that, for systemically significant financial actors, the combined operation of these two regimes together with the EU Charter of Fundamental Rights produces what it terms a regulatory trilemma: a three-cornered tension between (i) the AI Act's demand for transparency and human control, (ii) DORA's expectation of automation-capable operational resilience, and (iii) the Charter's guarantees of privacy, data protection, and non-discrimination. The article's contribution is threefold and deliberately modest. First, it brings the AI Act and DORA into a single doctrinal frame and shows that their interaction generates a genuine compliance dilemma rather than a relabelling of familiar bilateral tensions. Second, it situates that dilemma within the wider fragmentation of EU digital regulation, asking why DORA stands somewhat apart from the rights-and-innovation balance that De Gregorio and Dunn identify as the connecting thread of the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), and the AI Act, and whether the emerging discourse of digital sovereignty supplies a shared horizon. Third, it advances calibrated autonomy as a structured, proportionality-anchored interpretive and normative proposal that

conditions the permissible degree of AI autonomy on threat severity, data sensitivity, and decision reversibility. The article closes with reform proposals—joint guidance from the European Supervisory Authorities, a harmonised incident taxonomy from the European Union Agency for Cybersecurity (ENISA), a narrowly defined compliance safe harbour, and sector-specific advisory boards—assessed against questions of competence, legal basis, and feasibility. The analysis is confined to the financial sector governed by DORA; the closing section identifies which arguments travel beyond it.

Keywords: EU AI Act; DORA; cybersecurity; artificial intelligence; fundamental rights; regulatory trilemma; proportionality; calibrated autonomy; digital sovereignty; financial sector; operational resilience; data protection

1. Introduction

The deployment of AI in cybersecurity has moved from experimental application to operational practice. Financial institutions, critical-infrastructure operators, and public administrations across the European Union increasingly rely on AI-supported systems for threat detection, anomaly identification, automated incident response, and predictive vulnerability management [1,2]. The threat picture justifying this shift is sobering: the European Union Agency for Cybersecurity records a sustained rise in the volume and sophistication of attacks against EU entities, with the financial sector among the most heavily targeted [3], while widely cited industry estimates place the annual global cost of cybercrime in the trillions of euros and rising [4]. Such figures should be treated as orders of magnitude rather than precise measurements, but the direction of travel is not in doubt: the speed and scale of contemporary attack vectors—from polymorphic malware to AI-generated phishing—have outpaced the capacity of purely manual monitoring in large organisations.

The European Union's regulatory response to the twin challenges of AI governance and cybersecurity resilience has produced a normative landscape marked by internal tensions. The AI Act [5], adopted in 2024 as the first comprehensive horizontal AI regulation, imposes graduated obligations of transparency, explainability, and human oversight according to a system's risk classification [2]. DORA [6], in full application since 17 January 2025, requires financial entities to maintain robust Information and Communication Technology (ICT) risk-management frameworks capable of rapid, automation-supported detection and response [7]. Developed through distinct legislative processes and animated by different policy objectives, these regimes—read alongside the Charter—create, this article argues, a regulatory trilemma in AI-driven cybersecurity: a three-cornered tension between transparency, resilience, and fundamental rights that current interpretive guidance does not adequately resolve.

Contribution. It is important to state at the outset what is and is not claimed. This article does not assert that the trilemma is a newly discovered doctrinal category with binding legal force. It is, rather, an analytical framing that the author develops in order to render visible a compliance problem that the existing literature has tended to treat in pairwise terms. The contribution is threefold. First, the article brings the AI Act and DORA into one doctrinal frame and shows (in Section 4) that the simultaneous application of all three normative axes to a single system generates compliance dilemmas and interpretive conflicts that do not arise when any one axis is considered in isolation. Second, it locates that dilemma within the broader fragmentation of EU digital law (Section 5), asking why DORA appears less obviously animated by the innovation-and-rights balance that connects the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), and the AI Act, and whether the discourse of digital

sovereignty offers a common horizon. Third, it advances the principle of calibrated autonomy (Section 6) as a proportionality-anchored proposal, and it is careful to specify the legal status of that proposal and what it adds to ordinary proportionality reasoning.

Method and scope. The article proceeds by doctrinal legal analysis: it interprets the relevant provisions of the AI Act [5], DORA [6], the GDPR [8], and the Charter [9], reads them in light of their recitals and the surrounding scholarship, and identifies the points at which their requirements pull in different directions. The analysis is deliberately confined to the financial sector, because DORA is the operative resilience instrument there and supplies the concrete automation expectations on which the argument turns. Where the reasoning plausibly extends to other regulated sectors—for example, entities within the scope of the NIS2 Directive—this is noted; but the article does not claim that its conclusions hold uniformly across all domains of AI-driven cybersecurity, and Section 8 identifies the limits of generalisation.

The article proceeds as follows. Section 2 analyses the AI Act's risk classification as applied to cybersecurity systems and sets out the criteria that determine high-risk status. Section 3 examines DORA's ICT risk-management and incident-response requirements and the practical—as distinct from legal—pressure they exert toward AI-supported automation. Section 4 maps the normative conflicts and explains why a triadic framing is analytically distinct from three separate bilateral tensions. Section 5 develops the fundamental-rights dimension and situates the trilemma within the fragmentation of EU digital regulation and the emerging language of digital sovereignty. Section 6 considers why conventional norm-collision techniques are insufficient and then develops calibrated autonomy. Section 7 sets out reform proposals and tests them against competence, legal basis, and feasibility. Section 8 concludes and identifies the boundaries of the argument.

2. The AI Act's risk classification and cybersecurity AI systems

Regulation (EU) 2024/1689 establishes a four-tier risk taxonomy—unacceptable, high, limited, and minimal risk—that determines the obligations applicable to a given AI system [2,10]. For systems deployed in cybersecurity, the critical interpretive question is whether they fall within the high-risk category, which triggers the most onerous compliance duties, or are properly classified as limited- or minimal-risk systems subject only to transparency duties or none at all.

High-risk status under the AI Act is determined not by intuition or "functional analogy" but by two defined routes. The first is Article 6(1), which classifies as high-risk an AI system intended to be used as a safety component of a product, or itself a product, covered by the Union harmonisation legislation listed in Annex I, where that product must undergo third-party conformity assessment. The second is Article 6(2) read with Annex III, which enumerates specific high-risk use cases. Article 6(3), introduced during the trilogue, then carves out a significant exception: a system falling within an Annex III area is not high-risk where it does not pose a significant risk of harm to health, safety, or fundamental rights, including where it performs a narrow procedural task or merely improves the result of a previously completed human activity. The provider must document any such assessment. For cybersecurity systems, the analytical task is therefore to test the system against these specific criteria, not to reason by resemblance.

Applying that test, cybersecurity is not listed as a standalone high-risk domain in Annex III. Two Annex III categories may nonetheless capture particular cybersecurity systems. First, Annex III point 2

covers AI systems intended to be used as safety components in the management and operation of critical digital infrastructure; a cybersecurity system whose failure would directly endanger the operation of such infrastructure may fall within it. Second, Annex III point 1 covers certain biometric systems; a cybersecurity tool that performs behavioural-biometric authentication or identification may engage it, subject to the carve-outs and to the distinction between biometric identification and mere verification. A conventional network intrusion-detection system that processes traffic metadata without performing biometric identification, and whose output is reviewed before any consequential action, will frequently fall outside Annex III altogether, or within the Article 6(3) exception. The point for present purposes is that the high-risk question is genuinely indeterminate at the margins, and that this indeterminacy—rather than any blanket classification—is what generates legal uncertainty for deployers [10].

Where a cybersecurity system is high-risk, Article 14 requires effective human oversight, including measures enabling the natural persons to whom oversight is assigned to understand the system's capacities and limitations and to decide, in any particular situation, not to use it or to disregard, override, or reverse its output [11,12]. Article 13 requires that high-risk systems be sufficiently transparent to enable deployers to interpret and appropriately use their output. These duties, sensible for systems making consequential decisions about identified individuals, sit uneasily with cybersecurity systems that must act at machine speed against threats whose latency is measured in milliseconds. The tension is sharpened by Article 9, which requires risk-management processes addressing reasonably foreseeable risks to health, safety, and fundamental rights: in cybersecurity, the foreseeable harm of not acting autonomously—successful intrusion, data exfiltration, cascading failure—may itself be grave, and the AI Act offers no explicit metric for weighing that harm against the risks that human-oversight requirements are designed to contain.

3. DORA and the practical pressure toward automation-capable resilience

Regulation (EU) 2022/2554 establishes a comprehensive framework for the digital operational resilience of financial entities—credit institutions, investment firms, insurance undertakings, and, notably, ICT third-party service providers designated as critical. Its requirements became fully applicable on 17 January 2025 and span ICT risk management, incident classification and reporting, resilience testing, and third-party risk oversight [7].

Article 6 requires financial entities to maintain and review an ICT risk-management framework, including the identification, classification, and documentation of ICT-supported functions, processes, and assets, together with continuous monitoring. Article 10 requires mechanisms to promptly detect anomalous activities, with multiple layers of control and defined alert thresholds. Articles 17 to 23 impose a tightly timed incident-reporting regime, and Article 24 onward requires a digital operational resilience testing programme, including, for significant entities, threat-led penetration testing under Article 26 [7,13].

A careful claim about AI. It is essential to state precisely what DORA does and does not require. DORA does not, as a matter of law, mandate the use of artificial intelligence; its obligations are technology-neutral and framed in terms of capabilities—promptness, continuous monitoring, automated alerting, rapid classification—rather than particular techniques. The accurate claim is therefore a practical and conditional one: under present technological conditions, the combination of obligations DORA imposes on large, high-volume financial entities—real-time anomaly detection across very large

data flows, rapid containment before propagation, and classification within compressed reporting windows—creates strong operational pressure to deploy AI-supported systems, because human-only monitoring cannot reliably meet those expectations at that scale. This is a statement about technological tendency and the practical conditions of compliance, not about legal obligation. A smaller entity with simpler ICT estate may satisfy DORA without advanced AI; the pressure intensifies with scale, complexity, and transaction volume. The three concepts—what the law requires, what technology currently tends to make necessary, and what policy might favour—are kept distinct throughout this article, and the reader should read the term "automation-capable" in that light.

Two features of DORA reinforce this practical pressure without converting it into a legal mandate. First, the incident-reporting timelines—an initial notification, an intermediate report, and a final report on a compressed schedule following classification—favour systems capable of rapid, consistent triage; at high volumes, manual classification struggles to meet them reliably. Second, the obligation to detect and contain anomalies across complex, distributed environments rewards continuous automated monitoring. None of this is to say that DORA prescribes any particular architecture. It is to say that, for a Tier-1 institution processing very large transaction volumes, the path of least regulatory risk will, on current technology, frequently run through AI-supported tooling—and it is precisely that tooling which the AI Act may classify as high-risk.

4. Mapping the normative conflict

The conflict comes into focus when the obligations of the AI Act and DORA are mapped onto a single system—say, a machine-learning intrusion detection and response system (ML-IDRS) deployed by a systemically important bank and processing personal data in the course of monitoring.

First axis: autonomy *versus* oversight. DORA's emphasis on prompt detection and containment (Articles 10–11) presses the ML-IDRS toward acting—blocking suspicious traffic, isolating segments, suspending accounts—without awaiting human authorisation. If the system is high-risk, Article 14 of the AI Act requires that a human overseer be able to override or reverse its output. In a zero-day scenario where the attack window is measured in seconds, insisting on meaningful ex ante human authorisation before containment may render containment ineffective and so cut against DORA's resilience objective; yet permitting fully autonomous operation may be incompatible with the AI Act's oversight requirement. The conflict is real but not absolute: it is sharpest where the system is high-risk and the action is consequential and time-critical.

Second axis: transparency versus security. Article 13 requires sufficient transparency for deployers to interpret outputs, and Article 86 grants affected persons a right to an explanation of certain individual decisions [14,15]. In cybersecurity, however, disclosing detection methodologies, classification criteria, or response triggers may degrade effectiveness by handing adversaries a map for evasion. The cybersecurity instinct toward operational confidentiality is in tension with the AI Act's transparency paradigm [14]. The tension is manageable—transparency to a regulator or a vetted overseer differs from transparency to the data subject or the public—but it requires a calibrated answer rather than a categorical one.

Third axis: processing intensity versus data minimisation. Effective AI-driven cybersecurity processes large quantities of data—traffic metadata, system logs, behavioural patterns, and at times communication content. DORA's monitoring expectations presuppose substantial processing, while the GDPR's data-minimisation principle (Article 5(1)(c)) and the AI Act's data-governance duties for

high-risk systems (Article 10) constrain its volume and nature [16]. The intersection of cybersecurity necessity and data-protection restraint forms the third corner of the trilemma.

4.1. Why a trilemma, and not three bilateral tensions

A knowledgeable reader may object that each pairing—oversight versus automation, security versus transparency, monitoring versus minimisation—is already familiar, and that calling their conjunction a "trilemma" adds a label but no insight. The objection deserves a direct answer, because it identifies the precise point at which the contribution must be earned. The triadic framing matters for three reasons that a pairwise treatment cannot capture.

First, the three axes are not independent: a measure that relieves one tension typically aggravates another, so that the conflicts cannot be resolved one at a time. Increasing human oversight to satisfy the AI Act (axis one) slows containment and so strains DORA's resilience expectation; widening data collection to improve detection and meet DORA (axis three) deepens the privacy intrusion that the Charter constrains; narrowing disclosure to protect security (axis two) reduces the transparency the AI Act requires and weakens the evidentiary basis for demonstrating that a rights infringement was proportionate. The compliance problem is therefore genuinely three-bodied: each bilateral "solution" displaces the difficulty onto the third axis rather than dissolving it.

Second, the presence of the third axis alters the legal resolution of the other two. Were the question merely AI Act versus DORA, an interpreter might be tempted to resolve it by a lex specialis or sequencing argument—treating DORA as the more specific regime for financial entities and reading the AI Act's oversight duty down accordingly. But the fundamental-rights axis forecloses that shortcut: because Charter rights operate as a normative ceiling on both instruments, neither the AI Act nor DORA can be interpreted in a way that authorises a disproportionate rights infringement, regardless of which is treated as lex specialis. The third axis thus changes the available interpretive moves on the first two, rather than simply adding a further consideration to be balanced.

Third, the conjunction produces a compliance dilemma that is structurally distinct from the sum of its parts. An entity can, in principle, comply with any two axes at the cost of the third—maximal oversight and minimal data, but slow and so weak on resilience; or fast and data-rich, but weak on oversight and intrusive on rights. What it cannot do, under current guidance, is identify an authoritative point at which all three are simultaneously satisfied, because no instrument supplies the trade-off rule. It is this absence of a governing trade-off rule—not the mere coexistence of familiar tensions—that the trilemma names, and that the remainder of the article tries to address.

5. The fundamental-rights dimension and the fragmentation of EU digital regulation

The EU Charter of Fundamental Rights is the normative ceiling against which both the AI Act and DORA must be read. Three Charter rights are particularly engaged by AI-driven cybersecurity: respect for private and family life (Article 7), protection of personal data (Article 8), and non-discrimination (Article 21) [9].

AI-driven cybersecurity systems, and especially those employing user and entity behaviour analytics (UEBA), stand in tension with privacy by design rather than embodying it. Such systems construct behavioural profiles of employees, customers, and third parties, monitoring access, communication, and activity patterns to detect deviations indicative of compromise or insider threat. The

depth of profiling that effective detection tends to require—login and access patterns, data-transfer volumes, communication frequencies—constitutes a significant intrusion into the private sphere even when pursued for legitimate security purposes, and engages Articles 7 and 8 [3].

The non-discrimination dimension arises from the well-documented tendency of machine-learning systems to reproduce or amplify biases in their training data [17]. In cybersecurity this can manifest as systems that disproportionately flag the behaviour of employees from particular demographic groups, locations, or roles. A system that generates higher alert rates for access during non-standard hours—a pattern that may correlate with religious observance, caregiving responsibilities, or time-zone differences in global organisations—can produce indirectly discriminatory effects engaging Article 21.

How the Charter operates here should be stated with care, because an overly absolute formulation would be inconsistent with the proportionality method on which this article's own proposal rests. The Charter does not impose a flat prohibition, and it is not accurate to say that AI-driven cybersecurity "inevitably" infringes fundamental rights in a manner that admits no justification. Articles 7, 8, and 21 are not unqualified: under Article 52(1) of the Charter, limitations are permissible where they are provided for by law, respect the essence of the right, and are—subject to proportionality—necessary and genuinely meet objectives of general interest or the need to protect the rights of others. The more defensible claim is therefore the following: AI-driven cybersecurity, as ordinarily deployed, interferes with these rights, so that the operative question is not whether interference occurs but whether a given interference is lawful, necessary, proportionate, and accompanied by adequate safeguards. What distinguishes the rights axis from the AI Act–DORA conflict is not that it is absolute, but that it cannot be discharged purely by coordinating two secondary instruments: it requires an independent proportionality assessment that the coordination of the AI Act and DORA does not itself supply.

5.1. The trilemma as a symptom of regulatory fragmentation

The trilemma is best understood not as a drafting accident but as a symptom of how the EU has built its digital rulebook in successive, sector-specific layers. De Gregorio and Dunn argue that the GDPR, the Digital Services Act, and the AI Act, despite their different regulatory techniques, are connected by a single constitutional aspiration: the attempt to secure an optimal balance between innovation and the protection of fundamental rights, a balance they describe as the fil rouge of EU digital constitutionalism [1]. The AI Act fits this account comfortably. Although its architecture derives from risk-based product-safety regulation, it openly foregrounds fundamental rights, both in its stated objectives and in mechanisms such as the fundamental-rights impact assessment and the Article 86 right to explanation.

DORA sits less neatly within that thread. Its centre of gravity is operational resilience and financial-system stability; its conceptual lineage is closer to a security-and-continuity logic than to the rights-and-innovation balance that animates the GDPR–DSA–AI Act sequence. DORA contains comparatively little express engagement with the fundamental-rights consequences of the intensive monitoring it presupposes, leaving that engagement to the GDPR and now the AI Act. This is not a criticism of DORA on its own terms, but it explains why the trilemma arises: the rights-balancing work that is internal to the AI Act is, in DORA, largely external and assumed. When the two regimes meet in a single system, the financial entity is left to perform an integration that neither instrument fully performs for it.

5.2. Digital sovereignty as a possible common denominator

If DORA stands somewhat apart from the rights-and-innovation thread, is there a higher-order value under which resilience and rights can be reconciled? A candidate increasingly prominent in Brussels discourse is digital sovereignty—the Union's capacity to act autonomously in the digital domain and to set the terms on which digital infrastructure, data, and services that significantly affect European society are governed [18]. Read generously, digital sovereignty connects the two halves of the trilemma: it encompasses both the security and resilience of critical infrastructure (the concern that animates DORA) and the protection of citizens' data and rights and the projection of European values (the concern that animates the AI Act and the Charter).

The framing should be advanced with caution, for two reasons. First, digital sovereignty is a contested and under-defined concept; it is deployed inconsistently across EU documents and risks collapsing into industrial-policy capacity-building that marginalises the rights dimension rather than securing it. Second, invoking sovereignty does not by itself resolve any concrete conflict between an Article 14 oversight duty and an Article 11 containment expectation; it supplies a horizon, not a decision rule. Subject to those caveats, the concept is nonetheless useful here in a limited way: it identifies a shared objective—secure, rights-respecting European digital autonomy—against which the otherwise divergent logics of DORA and the AI Act can be read as complementary means rather than competing ends. That reframing is what makes a reconciling principle such as calibrated autonomy intelligible as something other than an ad hoc compromise.

6. Calibrated autonomy: a doctrinal proposal

6.1. Why conventional norm-collision techniques are insufficient

Before proposing a new principle, one should ask whether the existing legal toolkit already resolves the trilemma. Four conventional techniques present themselves, and each does part of the work without completing it.

Lex specialis derogat legi generali would treat DORA, as the specific regime for financial entities, as prevailing over the more general AI Act where they conflict. But the maxim is an imperfect fit: the AI Act expressly contemplates interaction with sectoral Union law rather than wholesale displacement, the two instruments largely regulate different objects (resilience outcomes versus AI-system properties), and—decisively—lex specialis cannot license a result that the Charter forbids. The risk-based classification of the AI Act itself does important work, since a system that is not high-risk escapes the Article 14 oversight duty altogether; but classification is precisely the question that is indeterminate at the margins (Section 2), and it does not tell a high-risk system how to reconcile oversight with containment. The general principle of proportionality is indispensable and underlies the proposal below, but in its unstructured form it offers little ex ante guidance to the engineer designing the system or the officer signing off on it. Ex post accountability—logging, audit, and review after the fact—is a necessary safeguard but, standing alone, accepts that the conflict will be resolved in the moment without a governing rule and reviewed only afterwards.

The conclusion is not that these techniques are wrong but that they are incomplete: each addresses one face of the problem and none supplies an operational trade-off rule that an entity can apply prospectively across all three axes. Calibrated autonomy is offered to fill that specific gap.

6.2. The principle and its legal status

Calibrated autonomy holds that the permissible degree of AI autonomy in cybersecurity operations should not be fixed by regulatory fiat but determined dynamically through a structured proportionality assessment turning on three variables: threat severity, data sensitivity, and decision reversibility.

Its legal status should be stated plainly, because the three possible registers carry different claims. Calibrated autonomy is advanced primarily as (a) an interpretive proposal—a reading of how the AI Act's oversight and transparency duties, DORA's resilience expectations, and the Charter's proportionality requirement can be made mutually consistent in their application to a single system—and, derivatively, as (b) a normative proposal addressed to regulators and the legislator, who could give it firmer footing through the guidance and reform measures set out in Section 7. It is not put forward as (c) a freestanding compliance heuristic that firms may treat as a safe harbour today; absent the supporting guidance, conformity with the framework does not guarantee conformity with the law. Distinguishing these registers matters: as an interpretive proposal it must be defensible as a plausible construction of the instruments as they stand; as a normative proposal it must be defensible as good policy; and it should not be read as already possessing the legal effect that only the reforms in Section 7 would confer.

What does calibrated autonomy add to ordinary proportionality, which already requires that any rights interference be suitable, necessary, and proportionate stricto sensu? Three things. First, it specifies the variables—threat severity, data sensitivity, reversibility—that the proportionality assessment must weigh in this particular context, converting an open-ended standard into a more determinate test. Second, it ties those variables to operational consequences through a tiered protocol (Section 6.4), so that the assessment can be embedded in system design ex ante rather than reconstructed by a court ex post. Third, by foregrounding reversibility it supplies a principled bridge to the AI Act's own concern—human oversight is demanded most strongly precisely where Article 14's capacity to "reverse" the output matters most, namely for irreversible actions. In short, calibrated autonomy is not a substitute for proportionality but a domain-specific structuring of it.

6.3. The three variables

Threat severity. Under conditions of acute threat—active ransomware propagation, a distributed denial-of-service attack on critical infrastructure, or an advanced persistent threat exfiltrating sensitive data—the balance shifts toward greater autonomous action. The proportionality calculus recognises that the harm prevented by immediate intervention (operational disruption, data loss, systemic contagion) may outweigh the harm of temporarily reduced human oversight. This is consistent with the necessity logic that underlies, for example, the GDPR's recognition of processing necessary to protect vital interests, and with proportionality as a general principle of EU law.

Data sensitivity. The intrusiveness of the system's processing should be calibrated to the data involved. Monitoring connection metadata and flow statistics is a lesser intrusion than deep-packet inspection of communication content. The principle requires defaulting to the least intrusive monitoring

sufficient for detection, escalating to more intensive processing only where specific indicators of compromise justify it. This operationalises the GDPR's data-minimisation principle and the AI Act's data-governance duties rather than displacing them.

Decision reversibility. This variable does the distinctive work. Blocking a suspicious IP address or quarantining a flagged file are readily reversible actions whose autonomous execution poses limited and remediable risk. Suspending a user account, terminating access credentials, or reporting a transaction as potentially fraudulent are less easily reversed and may cause significant and lasting harm to the individual. The principle requires meaningful human oversight for irreversible or high-consequence actions even under acute threat, while permitting full autonomy for reversible defensive measures. Reversibility thus maps directly onto Article 14's concern with the capacity to override or reverse, giving that concern operational content.

6.4. Operationalisation: a tiered protocol with a worked example

Operationally, calibrated autonomy would be embedded as a tiered response protocol. Under normal conditions (Tier 1) the system monitors and alerts, with consequential actions requiring human authorisation. Under elevated conditions (Tier 2) the system may autonomously execute reversible measures while flagging irreversible actions for human review. Under critical conditions (Tier 3) the system may act across the full range of defensive measures, subject to contemporaneous logging and post-hoc human review within a defined window. The operative tier is set algorithmically from predefined, auditable threat indicators, with human override available in either direction.

A single concrete scenario shows how the framework decides cases. Consider four candidate actions confronting the ML-IDRS, ordered by reversibility: (i) blocking suspicious inbound traffic from an IP range; (ii) quarantining a file flagged as malicious; (iii) suspending a user account showing signs of compromise; and (iv) reporting a customer transaction to authorities as potentially fraudulent. Actions (i) and (ii) are readily reversible and low-consequence; actions (iii) and (iv) are difficult to reverse and carry significant consequences for the individual—loss of access, reputational and possibly legal exposure. Table 1 shows how the protocol treats each action as the threat tier rises.

Table 1. Tiered response protocol under calibrated autonomy: treatment of candidate actions, ordered by reversibility, across threat tiers.

Action (by reversibility)	Tier 1—normal	Tier 2—elevated	Tier 3—critical
Block suspicious IP (reversible)	Alert; human authorises	Autonomous	Autonomous
Quarantine flagged file (reversible)	Alert; human authorises	Autonomous	Autonomous
Suspend user account (hard to reverse)	Human authorisation	Flagged for human review	Autonomous, with immediate logging and prompt post-hoc review
Report transaction as fraudulent (hard to reverse)	Human authorisation	Human authorisation	Autonomous only where law permits, with immediate logging and prompt review

The example illustrates the framework's logic: autonomy expands with threat severity but is constrained by reversibility, so that even at the critical tier the most consequential and least reversible action (reporting an individual) remains the most tightly controlled and is permitted autonomously only where the applicable law independently allows it. The tier governs how much autonomy the severity of the threat warrants; reversibility governs which actions that autonomy may reach; and data sensitivity governs how intrusively the system may gather the evidence on which it acts.

7. Regulatory reform proposals

Calibrated autonomy requires institutional support to move from interpretive proposal to operative practice. This section sets out four measures and tests each against three questions the reviewers rightly press: which body has competence to act, on what legal basis, and whether the measure is feasible and free of serious downside.

7.1. Joint guidance from the European Supervisory Authorities

The European Supervisory Authorities (ESAs)—in particular the European Banking Authority and the European Securities and Markets Authority—should issue joint guidelines on applying the AI Act to cybersecurity systems in the financial sector, addressing the classification question, the interaction of DORA's automation expectations with the AI Act's oversight duties, and the operationalisation of calibrated autonomy. As to competence and legal basis, the ESAs may issue guidelines and recommendations under Article 16 of their founding Regulations, and they already exercise extensive joint rule-making under DORA itself, which provides a natural vehicle. Such guidelines are formally non-binding but operate through a "comply-or-explain" mechanism that gives them substantial practical force. The principal downside is the limit of competence: the ESAs cannot authoritatively interpret the AI Act, whose governance is allocated to the AI Office and national market-surveillance authorities, so guidance would need to be coordinated with those bodies to avoid divergent readings. This is a reason for inter-authority cooperation, not an obstacle to it.

7.2. A harmonised incident taxonomy from ENISA

The European Union Agency for Cybersecurity (ENISA) should develop a harmonised incident-classification taxonomy bridging DORA's incident categories and the AI Act's risk levels, reducing the burden of operating under dual classifications and supporting consistent calibration. ENISA's competence to develop such technical instruments and to support coherent implementation across Member States is grounded in the Cybersecurity Act, and its existing role in incident-reporting frameworks under NIS2 makes the task institutionally plausible. The main feasibility concern is coordination cost—any taxonomy must be reconciled with the implementing and regulatory technical standards already adopted under DORA—so the measure is best framed as harmonisation of existing categories rather than the creation of a competing scheme.

7.3. A narrowly defined compliance safe harbour

Of the four proposals this requires the closest scrutiny, because a loosely drawn safe harbour could license rights infringements by fiat. The proposal is therefore deliberately narrow. It would provide that an entity which deploys an AI cybersecurity system in conformity with the calibrated-autonomy framework, and which maintains adequate documentation, conducts post-incident review, and remediates any identified rights infringement, benefits from a rebuttable presumption that, in taking autonomous reversible defensive measures during a genuine acute threat, it satisfied the AI Act's oversight requirement and DORA's resilience requirement.

Three limits define what the safe harbour does and does not do. First, as to subject matter, it would attach only to reversible defensive measures taken under verified acute-threat conditions; it would not shield irreversible or high-consequence actions, which remain subject to human oversight. Second, as to legal effect, it would create a rebuttable evidential presumption of compliance, not immunity; it could not and would not displace the Charter, the GDPR, or any data subject's rights, and a demonstrated disproportionate infringement would rebut it. Third, as to legal basis, a safe harbour with genuine legal effect cannot be conjured by guidance alone: it would require a legislative footing, most plausibly through a targeted amendment to, or delegated act under, the relevant instrument, and its design would have to respect the limits the Charter places on any limitation of rights. Framed in this confined way, the safe harbour addresses the chilling effect of regulatory uncertainty on legitimate defensive automation without becoming a charter for unaccountable action; framed more broadly, it should be rejected.

7.4. Sector-specific advisory boards

The EU could encourage, within critical sectors, AI-cybersecurity advisory boards composed of security professionals, data-protection experts, and civil-society representatives, to conduct prospective impact assessments, review tier-escalation protocols, and issue sector-specific guidance on proportionate processing. The honest assessment is that this is the softest of the four proposals: its value is deliberative rather than legal, and it risks adding a layer of process without clear authority. It is therefore proposed not as a mandate but as a recommended governance practice that entities and supervisors may adopt, complementing—rather than substituting for—the existing fundamental-rights impact assessment under the AI Act and the data-protection impact assessment under the GDPR. So framed, it institutionalises multidisciplinary deliberation that current frameworks leave to individual organisations, while avoiding the creation of a new and ill-defined locus of competence.

8. Conclusion

The deployment of AI in financial-sector cybersecurity poses a regulatory challenge that cuts across AI governance, financial regulation, and fundamental-rights protection. The regulatory trilemma identified here—the three-cornered tension between transparency, resilience, and rights—is not merely an incidental drafting imperfection but a reflection of how the EU has legislated its digital domain in successive, differently motivated layers, with DORA's resilience logic standing somewhat apart from the rights-and-innovation thread that connects the GDPR, the DSA, and the AI Act.

Rather than restate the urgency of the problem, it is worth fixing two concrete legal findings that the analysis yields. First, the sharpest conflict is narrower than a broad reading might suggest: it bites only where a cybersecurity system is properly classified as high-risk under the AI Act and is called upon to take a consequential, time-critical, and difficult-to-reverse action; for the large class of reversible defensive measures, and for systems outside the high-risk category, the AI Act and DORA can be reconciled without strain. Second, the fundamental-rights axis does not function as an absolute bar but as a proportionality constraint under Article 52(1) of the Charter; its distinctive force is that it cannot be discharged by coordinating the AI Act and DORA alone, but requires an independent necessity-and-proportionality assessment that calibrated autonomy is designed to structure.

Calibrated autonomy does not dissolve these tensions; it offers a disciplined way of managing them by anchoring the permissible degree of AI autonomy in a structured proportionality assessment—threat severity, data sensitivity, reversibility—operationalised through tiered protocols and supported by guidance, a harmonised taxonomy, a narrowly drawn safe harbour, and advisory deliberation. The argument has been confined to the financial sector, where DORA supplies the operative resilience expectations. Its first axis—oversight versus automation—and its rights analysis travel readily to other regulated settings, such as essential-service operators within the scope of NIS2, where comparable automation pressures arise; its DORA-specific elements, including the precise reporting timelines and the safe-harbour design, do not transpose automatically and would require separate analysis. Within those limits, the EU, having assumed a leading role in both AI and cybersecurity regulation, has a corresponding responsibility to ensure that its frameworks operate in concert rather than at cross-purposes.

Declaration of generative AI and AI-assisted technologies

The authors did not use generative AI or AI-assisted technologies in the writing of this manuscript. The authors take full responsibility for the content of the manuscript.

Authors' contribution

Conceptualization, F.T.; methodology, F.T.; formal analysis, F.T.; investigation, F.T.; resources, F.T.; writing—original draft preparation, F.T.; writing—review and editing, F.T. and B.S.; supervision, B.S.; project administration, F.T. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] De Gregorio G, Dunn P. The European risk-based approaches: connecting constitutional dots in the digital age. *Common Mark. Law Rev.* 2022, 59(2):473–500.

[2] Cancela-Outeda S. The EU's AI act: a framework for collaborative governance. *Internet Things* 2024, 27:101291.

[3] European Union Agency for Cybersecurity (ENISA). ENISA threat landscape 2024. 2024. Available: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024 (accessed on 6 July 2026).

[4] Cybersecurity Ventures. Official cybercrime report (annual cost projection). 2025. Available: https://cybersecurityventures.com/official-cybercrime-report-2025/ (accessed on 6 July 2026).

[5] 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 and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). 2024. Available: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (accessed on 6 July 2026).

[6] European Union. Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014, (EU) No 909/2014 and (EU) 2016/1011 (Text with EEA relevance). 2022. Available: https://eur-lex.europa.eu/eli/reg/2022/2554/oj/eng (accessed on 6 July 2026).

[7] Novelli C, Casolari F, Rotolo A, Taddeo M, Floridi L. Taking AI risks seriously: a new assessment model for the AI act. *AI Soc*. 2024, 39(5):2493–2497.

[8] 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, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). 2016. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng (accessed on 6 July 2026).

[9] European Union. Charter of fundamental rights of the European union. 2012. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=oj:JOC_2012_326_R_0391_01 (accessed on 6 July 2026).

[10] Veale M, Borgesius FZ. Demystifying the draft EU artificial intelligence act. *Comput. Law Rev. Int*. 2021, 22(4):97–112.

[11] Martini M, Wendehorst C. *KI-VO: Verordnung über Künstliche Intelligenz: Kommentar*, 2nd ed. Munich: C. H. Beck, 2024. pp. 1178.

[12] Panezi M. Human oversight. In *The EU Artificial Intelligence (AI) Act: A Commentary*. Alphen aan den Rijn: Kluwer Law International B.V., 2024.

[13] Buttigieg C, Zimmermann C. The digital operational resilience act: challenges and some reflections on the adequacy of Europe's architecture for financial supervision. *ERA Forum* 2024, 25(1):11–28.

[14] Busuioc M, Curtin D, Almada M. Reclaiming transparency: contesting the logics of secrecy within the AI Act. *Eur. Law Open* 2023, 2(1):79–105.

[15] Kaminski ME, Malgieri G. The right to explanation in the AI act. *SSRN* 2025.

[16] Van Bekkum M. Using sensitive data to de-bias AI systems: article 10(5) of the EU AI act. *Comput. Law Secur. Rev*. 2025, 56:106115.

[17] Yordanova K. Data and data governance. In *The EU Artificial Intelligence (AI) Act: A Commentary*. Alphen aan den Rijn: Kluwer Law International B.V., 2024.

[18] Madiega TA. Digital sovereignty for Europe. 2020. Available: https://www.europarl.europa.eu/RegData/etudes/BRIE/2020/651992/EPRS_BRI(2020)651992_EN.pdf (accessed on 6 July 2026).