unesco

PATHWAYS ON CAPACITY BUILDING FOR AI SUPERVISORY AUTHORITIES

INSIGHTS AND RECOMMENDATIONS FROM THE 1ST UNESCO EXPERT ROUNDTABLE ON AI SUPERVISION

Funded by
the European Union

Published in 2025 by the United Nations Educational, Scientific and Cultural Organization,
7, place de Fontenoy, 75352 Paris 07 SP, France

SHS/REI/EAI/CAAI/2025/PR

This document is produced in the context of the project "Supervising AI by Competent Authorities" in implementing UNESCO's Recommendation on the Ethics of AI through capacity building in partnership with the Dutch Authority for Digital Infrastructure (RDI). This project is financed by the European Union via the Technical Support Instrument (TSI). UNESCO is a knowledge partner in the context of this project.

This document was produced with the financial assistance of the European Union. The views expressed herein can in no way be taken to reflect the official opinion of the European Union.

Editors: Dr. Doaa Abu Elyounes, Dr. Yannic Duller, Dr. Angelica Fernandez, Mirela Kmetic-Marceau

Graphic design: UNESCO

Printed by: UNESCO

Printed in Paris

unesco

PATHWAYS ON CAPACITY BUILDING FOR AI SUPERVISORY AUTHORITIES

INSIGHTS AND RECOMMENDATIONS FROM THE 1ST UNESCO EXPERT ROUNDTABLE ON AI SUPERVISION

Dutch Authority for Digital Infrastructure
Ministry of Economic Affairs and Climate Policy

Funded by the European Union

CONTENTS

INTRODUCTION

THE EXPERT ROUNDTABLE ON CAPACITY BUILDING FOR AI SUPERVISORY AUTHORITIES

This report grew out of the momentum created by the first Expert Roundtable on Capacity Building for AI Supervisory Authorities, held at UNESCO Headquarters in May 2025. The Roundtable brought together supervisors, policymakers, and researchers who share a common concern: how to equip supervisory authorities with the tools and knowledge needed to govern artificial intelligence (AI) effectively. The discussions revealed both the diversity of national experiences and a remarkable convergence around key challenges such as building institutional capacity, fostering regulatory cooperation, and translating high-level principles into operational practice.

The present volume continues that exchange in written form. It gathers the contributions of several experts who participated in the Roundtable and invites readers to engage with their analyses in greater depth. It is conceived as a public resource offering structured knowledge that can inform supervisory practices well beyond the

event itself. Each contribution reflects the author's independent analysis and field experience, illustrating how concepts such as "responsible innovation," "institutional experimentalism," "interpretative supervision," or "trust-based regulation" can be translated into practical approaches for authorities. The diversity of perspectives underscores that building supervisory capacity is not a purely technical exercise but an institutional transformation that requires openness, cooperation, and sustained learning.

The report forms part of the project Supervising AI by Competent Authorities, a joint initiative of the Dutch Authority for Digital Infrastructure (RDI) and UNESCO, funded by the European Commission's SG REFORM, which supports national and regional authorities in developing the competences required to supervise AI under the EU AI Act and related governance frameworks. Within this context, the report offers an additional layer of insight: it distils lessons from practitioners who have worked in the front lines of supervision, translating legal obligations and ethical principles into day-to-day institutional routines.

The value of this collection lies in its pragmatism. The authors write from experience: they have designed sandboxes, coordinated cybersecurity frameworks, advised supervisors, and studied institutional change from within. Their insights shed light on how supervision can evolve from a reactive control function into a proactive mechanism that enables responsible innovation. For instance, contributions examine how regulatory sandboxes can foster trust when designed with transparency and inclusiveness; how supervisory authorities can overcome siloed governance structures through cross-sectoral coordination; how guardrails such as incident reporting and "red lines" can turn oversight into a driver of safety and accountability; or how interpretative supervision can help supervisors make sense of complex, adaptive AI systems when traditional audit methods fall short.

Taken together, the chapters illustrate a shared conviction: that supervision and innovation are not opposing forces but mutually reinforcing pillars of a trustworthy digital future. Effective supervision requires technical competence, but equally a capacity for interpretation, dialogue, and anticipation. It demands institutions that learn continuously, cooperate across sectors and borders, and translate ethical and legal norms into operational practice.

By offering concrete reflections and methodological guidance, this report aims to support supervisory authorities as they consolidate their mandates under the EU AI Act and parallel frameworks worldwide. It is intended as a practical reference for officials, policymakers, and experts seeking to design agile oversight structures that protect the public interest while allowing beneficial technologies to flourish. In doing so, it contributes to a growing international conversation, anchored in the UNESCO Recommendation on the Ethics of Artificial Intelligence, about how to ensure that AI governance remains both effective and human-centred.

SECTION I.

IMPLEMENTATION MODELS & STRATEGIES FOR COMPETENT AUTHORITIES

Supervisory authorities around the world are entering a defining phase of institutional transformation. As artificial intelligence becomes embedded in essential functions of society, from healthcare and social protection to financial markets and public information, the question is no longer whether authorities should adapt, but how. The first panel of UNESCO's Expert Roundtable on Capacity Building for AI Supervisory Authorities addressed precisely this issue: what does it take to translate the abstract mandates of AI regulation into operational institutions capable of governing learning, adaptive technologies?

The discussion highlighted that implementation is as much about culture and capability as it is about structure. Participants emphasised that authorities need more than new legal powers; they require new reflexes. Peer-to-peer learning networks such as the European Working Group of Competent Authorities on AI and the NOBAREG network have shown how shared experience can accelerate institutional readiness. This insight underpins initiatives like the UNESCO Global Network of AI Supervisory Authorities, designed to provide the "right people in the right room at the right time" to foster dialogue, exchange, and joint problem-solving. The session also pointed to the growing relevance of SUPTech – the use of technology by supervisors themselves – as an emerging frontier of AI supervision, capable of enabling real-time monitoring and secure data exchange between authorities and developers.

At the heart of these debates lies a common recognition: effective supervision of AI requires new institutional models that combine technical insight, regulatory judgment, and anticipatory intelligence. Competent authorities must be able not only to enforce rules but to interpret behaviour, identify emerging risks, and learn collectively from evidence. This evolution calls for governance architectures that embed learning, monitoring, and coordination into the very fabric of supervision.

Kevin Zandermann's contribution, The OBSERVE Framework: A Global Scaling of Interpretative Supervision, elaborates precisely such an architecture. Building on the idea that AI supervision is no longer about static compliance but about continuous interpretation, it proposes a structured model for developing institutional intelligibility. The OBSERVE framework sets out seven interlinked pillars, from observatory units and behavioural monitoring to evidence libraries and early-warning systems, that together enable authorities to detect, understand, and respond to the dynamic behaviour of AI systems. It captures the same spirit that animated the roundtable discussions: supervision as an evolving practice rooted in cooperation, foresight, and the intelligent use of data.

THE OBSERVE FRAMEWORK: SCALING INTERPRETATIVE SUPERVISION INTO A GLOBAL ARCHITECTURE FOR AI GOVERNANCE

Author: Kevin Zandermann; Tony Blair Institute

Abstract

This paper introduces the OBSERVE framework, a comprehensive institutional model designed to build supervisory capacity for governing adaptive and increasingly autonomous AI systems. Recognising that AI supervision now requires continuous interpretation rather than static compliance checking, the framework operationalises seven interconnected components: Observatory Units, Behavioural Monitoring, Stakeholder Networks, Evidence Libraries, Responsive Frameworks, Values Integration, and Early Warning Systems. Together, these elements enable authorities to make AI system behaviour intelligible, contextualised, and actionable across diverse real-world settings. The contribution argues that the shift toward interpretative supervision is essential in a technological landscape marked by opacity, emergent behaviours, and evolving system capabilities. By offering a scalable and adaptable architecture, the OBSERVE framework provides policymakers and supervisory authorities with a structured pathway toward anticipatory, context-sensitive, and democratically accountable governance of AI systems at national and global levels.

Executive Summary

Artificial intelligence (AI) is increasingly mediating decisions once firmly grounded in human judgment and public accountability: a diverse array of systems ranging from machine-learning classifiers and predictive analytics to recommender engines, large-scale generative models, and autonomous AI agents shape how creditworthiness is assessed,[1] how patients are triaged for treatment,[2] how resources are allocated, and how citizens encounter information in the public sphere[3]. The expanding influence of these systems, intensified by the rapid diffusion of large language models, has prompted calls for credible and effective mechanisms of supervision.[4]

In developing such mechanisms, policymakers have often looked at established governance models in sectors such as aviation and nuclear safety, domains where complex technologies have been rendered governable through codified standards, measurable safety goals, and traceable engineering processes.[5] However, the transferability of such models to AI governance is limited as they rest on the premise that systems can be exhaustively specified, their causal chains mapped, and their risks contained through design and oversight.[6]

By contrast, AI systems do not operate as closed mechanical infrastructures but as adaptive processes,[7] whose behaviour emerges from statistical inference and feedback dynamics rather than from fixed engineering parameters. They evolve continuously with new data and interactions, and their architectures blur traditional boundaries between human and machine agency.[8] The significance of their outputs, whether social, ethical, or legal, cannot be directly inferred from code or documentation but must be interpreted within specific contexts of use.

Established ethical frameworks, such as bioethics, also provide only partial traction for AI governance, as they presuppose identifiable agents, discrete harm, and stable contexts of intervention.[9] AI systems, by contrast, operate across domains and generate emergent, systemic effects with dispersed agency and accountability shaped as much by data infrastructures as by human intentions.

[1] Wilhelmina Afua Addy and others, 'AI in Credit Scoring: A Comprehensive Review of Models and Predictive Analytics' (2024) 18 Global Journal of Engineering and Technology Advances 118.
[2] Cansu Yüksel Elgin and Ceyhun Elgin, 'Ethical Implications of AI-Driven Clinical Decision Support Systems on Healthcare Resource Allocation: A Qualitative Study of Healthcare Professionals' Perspectives' (2024) 25 BMC Medical Ethics 148.
[3] Aaron Hyzen and others, 'Epistemic Welfare and Algorithmic Recommender Systems: Overcoming the Epistemic Crisis in the Digitalized Public Sphere' [2025] Communication Theory qtaf018.
[4] Chinasa T Okolo, 'Governance of AI-Based Algorithms', Handbook of Human-Centered Artificial Intelligence (Springer, Singapore 2025) <https://link.springer.com/rwe/10.1007/978-981-97-8440-0_84-1> accessed 12 November 2025.
[5] Brian Judge, Mark Nitzberg and Stuart Russell, 'When Code Isn't Law: Rethinking Regulation for Artificial Intelligence' (2025) 44 Policy and Society 85.
[6] Amna Batool, Didar Zowghi and Muneera Bano, 'AI Governance: A Systematic Literature Review' (2025) 5 AI and Ethics 3265.
[7] Marijn Janssen, 'Responsible Governance of Generative AI: Conceptualizing GenAI as Complex Adaptive Systems' (2025) 44 Policy and Society 38.
[8] Dino Pedreschi and others, 'Human-AI Coevolution' (arXiv, 6 May 2024) <http://arxiv.org/abs/2306.13723> accessed 12 November 2025.
[9] Luciano Floridi (ed), Ethics, Governance, and Policies in Artificial Intelligence, vol 144 (Springer International Publishing 2021) <https://link.springer.com/10.1007/978-3-030-81907-1> accessed 12 November 2025.

In short, AI introduces a moving target for regulation: a technology that learns, generalises, and transforms alongside the societies that deploy it.[10] AI Supervisory authorities thus face not only a problem of control but a problem of meaning, as they seek to make the behaviour of AI systems intelligible within moral, legal and social frameworks never designed for learning, autonomous and continuously evolving technologies. Addressing this interpretive gap requires institutions capable of translating the behaviour of complex systems into forms of understanding that support accountability, legitimacy, and effective governance.

This paper advances interpretative supervision as a model for governing AI systems. Interpretative supervision denotes the institutional capacity to understand, contextualise, and act upon AI system behaviour, translating technical outputs into legally meaningful, ethically grounded, and socially intelligible judgments. It builds on the insights of XAI 2.0, which reframes explainability as a pluralistic, context-dependent, and action-oriented process rather than a purely technical property of AI models.[11]

Interpretative supervision anchors oversight in contextual understanding, adaptive response, and anticipatory foresight. Its goal is not to decode every model parameter, but to cultivate institutions capable of situating AI behaviour within the frameworks of accountability, human rights, and public trust. In this sense, interpretative supervision operationalises the shift from technical interpretability to institutional intelligibility, ensuring that societies can govern learning systems not through perfect knowledge, but through sustained interpretative capacity.

To operationalise this vision, the paper proposes the OBSERVE framework, a blueprint for institutional capacity-building that enables authorities to detect emerging harms, anticipate systemic risks, and align AI development with democratic values and public safety. OBSERVE comprises seven interdependent elements:

- ***O****bservatory units* that serve as specialised monitoring and interpretation hubs within regulatory agencies.
- ***B****ehavioural monitoring systems* that provide continuous, real-time oversight rather than episodic audits.
- ***S****takeholder networks* that connect supervisors with domain experts, civil society, and affected communities.
- ***E****vidence libraries* that preserve institutional memory of precedents, failures, and corrective actions.

[10] Noam Kolt, Michal Shur-Ofry and Reuven Cohen, 'Lessons from Complex Systems Science for AI Governance' (2025) 6 Patterns 101341.
[11] Luca Longo and others, 'Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions' (2024) 106 Information Fusion 102301.

- ***R**esponsive frameworks* that replace rigid compliance checklists with adaptive, outcome-oriented regulation.
- ***V**alue integration mechanisms* that embed fairness, accountability, and proportionality into oversight practice.
- ***E**arly-warning systems* for horizon scanning, scenario modelling, and cross-border intelligence sharing.

Certain components of this framework, particularly evidence libraries and responsive frameworks, are informed by the supervision of complex and nonlinear systems such as financial markets and public health. In those domains, supervisory authorities recognised that uncertainty and interdependence preclude purely rule-based control, and instead developed institutions oriented toward learning, feedback, and adaptive management. [12] These traditions offer valuable insights for AI supervision, where effectiveness is likely to be achieved not through predictability but through the capacity to interpret and respond dynamically to evolving risks. Implementation of the framework can proceed incrementally. Emerging supervisors may begin by designating oversight leads and developing structured case libraries. Mid-capacity authorities can establish observatory units and stakeholder networks to enhance interpretive reach. Advanced jurisdictions can deploy real-time supervisory technologies and lead international coordination.

Why AI Defies Established Governance Models

Artificial intelligence presents regulatory challenges that defy the successful precedents set by aviation, nuclear power, or electricity.[13] Whereas these technologies could be largely governed through clear technical standards, well-defined safety goals, and traceable engineering processes, AI resists such treatment. It is a general-purpose technology whose applications cut across every sector of the economy, but unlike electricity or the Internet, it lacks standardised definitions and protocols that might anchor regulation.[14]

The rise of deep learning architectures amplifies this difficulty. These AI systems do not operate as machines with parts that can be inspected or certified; their behaviours emerge from complex, data-driven training processes that are often opaque even to their developers.[15] The resulting outputs are probabilistic and context-dependent, producing behaviours that cannot be fully predicted or guaranteed to conform to predefined specifications.

[12] World Health Organization, Systems Thinking: For Health Systems Strengthening (2009).
[13] Brian Judge, Mark Nitzberg and Stuart Russell, 'When Code Isn't Law: Rethinking Regulation for Artificial Intelligence' (2025) 44 Policy and Society 85.
[14] Brian Judge, Mark Nitzberg and Stuart Russell, 'When Code Isn't Law: Rethinking Regulation for Artificial Intelligence' (2025) 44 Policy and Society 85.
[15] Maarten Goos and Maria Savona, 'The Governance of Artificial Intelligence: Harnessing Opportunities and Mitigating Challenges' (2024) 53 Research Policy 104928.

This technical opacity is reinforced by the structure of AI's innovation ecosystem. Frontier development is led largely by private laboratories and open-source communities rather than by public research institutions, leaving supervisory authorities positioned between concentrated market power on one side and the rapid, decentralised diffusion of increasingly capable systems on the other.[16] Within this landscape, even the notion of safety becomes elusive. Preventing a plane crash or a reactor failure is a finite technical objective, while aligning adaptive systems with human values is an open, contested, and continuously evolving endeavour that resists codification.[17]

The governance challenge is further amplified by the constraints of existing ethical frameworks. Traditions such as bioethics, which institutionalised principles of consent, beneficence, and non-maleficence, presuppose a human agent, an identifiable subject of harm, and a stable context of intervention.[18] AI systems violate these assumptions: they operate across domains, generate cumulative and distributed effects, and produce consequences that are systemic rather than individual. Agency is diffused across algorithms, datasets, developers, and deployers, producing accountability gaps that challenge conventional notions of responsibility and liability.

These structural and ethical discontinuities reveal that the governance of AI cannot be reduced to the management of technical risks or the extension of existing supervisory templates. Aligning adaptive systems with human values is not merely a procedural question of compliance, but a deeper philosophical and institutional challenge. It demands new modes of supervision capable of interpreting how learning AI systems act, evolve, and interact with human institutions over time.

This growing challenge has produced a divergence in global approaches to AI governance. The European Union's AI Act establishes a comprehensive, risk-based framework extending oversight across the AI lifecycle.[19] The United States adopts a model grounded in voluntary industry standardisation through the Department of Commerce's NIST AI Risk Management Framework.[20] China, by contrast, embeds AI governance within a broader system of state oversight, national security, societal cohesion, and ideological stability.[21]

Each reflects distinct institutional logics: the emphasis on the precautionary principle[22] and fundamental rights of the EU, the market-driven standardisation of the US,[23] and China's state-centred coordination of

[16] Araz Taeihagh, 'Governance of Generative AI' (2025) 44 Policy and Society 1.
[17] Iason Gabriel and Vafa Ghazavi, 'The Challenge of Value Alignment: From Fairer Algorithms to AI Safety' (arXiv, 19 January 2021) <http://arxiv.org/abs/2101.06060> accessed 12 November 2025.
[18] Elizabeth Seger, 'In Defence of Principlism in AI Ethics and Governance' (2022) 35 Philosophy & Technology 45
[19] 'EU AI Act: First Regulation on Artificial Intelligence' (Topics | European Parliament, 8 June 2023) <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence> accessed 12 November 2025.
[20] 'AI Risk Management Framework' (NIST, 12 July 2021) <https://www.nist.gov/itl/ai-risk-management-framework> accessed 12 November 2025.
[21] Liza Mark and Tianyun (Joyce) Ji, 'China Publishes the AI Security Governance Framework' (Haynes and Boone 2024).
[22] 'Precautionary Principle - EUR-Lex' <https://eur-lex.europa.eu/EN/legal-content/glossary/precautionary-principle.html> accessed 12 November 2025.
[23] 'NIST Strategy for American Technology Leadership in the 21st Century' (NIST, 2 September 2025) <https://www.nist.gov/director/strategic-priorities> accessed 12 November 2025.

technology, security, and social order.[24] Yet all face the same structural dilemma: how to govern a technology that learns, adapts, and evolves faster than the institutions that are designed to supervise it.

The Mechanistic Interpretability Pathway

For more than a decade, researchers pursued what seemed an obvious solution to the governance of AI: mechanistic interpretability. The goal was to reverse-engineer neural networks neuron by neuron, mapping each unit to a discrete function until models were as legible as circuit diagrams. Many seemingly promising solutions encouraged this path: feature visualisation suggested neurons acted as edge or object detectors, saliency maps appeared to highlight the input features driving decisions, and sparse autoencoders compressed neural activations into seemingly interpretable codes.[25]

But as the field matured, these techniques exposed their own illusions. Hendrycks and Hiscott's influential 2025 critique[26] to Dario Amodei's ambition of achieving an "MRI for AI,"[27] documents the shortfalls of these solutions: saliency maps applied to random, untrained networks, produced heatmaps nearly indistinguishable from trained ones, showing they revealed mathematical quirks of architectures rather than reasoning, while sparse autoencoders underperformed simple baselines, with DeepMind deprioritising this year's line of research, conceding it had produced "limited actionable insights.[28]

However, the deeper shortfall was to treat mechanistic interpretability as the central foundation for solid regulation. Just as meteorologists cannot predict storms by tracking molecules, and neuroscientists cannot explain consciousness by cataloguing synapses, supervisors cannot hope to govern AI by atomistic reduction.

For supervisors, this recognition should be liberating rather than limiting. It suggests that the goal should not be to achieve total transparency of AI systems, an impossible standard that has the potential to paralyse regulatory action, but to develop institutional capacity for interpreting AI behaviours at the level where they intersect with meaningful legal, ethical, and social concerns. This perspective resonates with UNESCO's broader shift toward capacity-building in AI governance: moving beyond the technical aspiration to fully decode models toward the practical task of developing institutional competence, regulatory learning, and cross-border collaboration.

[24] Jinghan Zeng, Artificial Intelligence with Chinese Characteristics: National Strategy, Security and Authoritarian Governance (Springer 2022) <https://link.springer.com/10.1007/978-981-19-0722-7> accessed 12 November 2025.
[25] 'The Misguided Quest for Mechanistic AI Interpretability' (AI Frontiers) <https://ai-frontiers.org/articles/the-misguided-quest-for-mechanistic-ai-interpretability> accessed 12 November 2025.
[26] 'The Misguided Quest for Mechanistic AI Interpretability' (AI Frontiers) <https://ai-frontiers.org/articles/the-misguided-quest-for-mechanistic-ai-interpretability> accessed 12 November 2025.
[27] 'Dario Amodei — The Urgency of Interpretability' <https://www.darioamodei.com/post/the-urgency-of-interpretability> accessed 12 November 2025.
[28] DeepMind Safety Research, 'Negative Results for Sparse Autoencoders On Downstream Tasks and Deprioritising SAE Research' (Medium, 26 March 2025) <https://deepmindsafetyresearch.medium.com/negative-results-for-sparse-autoencoders-on-downstream-tasks-and-deprioritising-sae-research-6cadcfc125b9> accessed 12 November 2025.

XAI 2.0 as the Foundation of Interpretative Supervision

A significant part of the technical AI community has begun to pivot from mechanistic interpretability toward more pragmatic and interdisciplinary approaches. This transition, often described as the shift from "XAI 1.0" to "XAI 2.0," reflects hard-won lessons about what makes explanations useful in practice. The decisive statement of this turn came in the 2024 XAI 2.0 manifesto published in Information Fusion, where leading researchers identified 28 open problems across nine categories.[29]

The manifesto's core insight is that explanations must be pluralistic. Different stakeholders, such as developers debugging models, supervisors ensuring compliance, clinicians making treatment decisions, and citizens challenging algorithmic determinations, require different types of explanations at different levels of abstraction. No single technical method can serve all interpretative needs. Effective explanations, the manifesto argues, must be:

1. **Context-based:** Grounded in human-meaningful categories rather than raw statistical features. Instead of highlighting pixels, explanations should reference medically relevant patterns, legally protected characteristics, or ethically significant concepts.

2. **Falsifiable:** Capable of supporting counterfactual reasoning and empirical testing. Explanations should enable supervisors to ask: "What would happen if...," and receive testable predictions.

3. **Actionable:** Designed to support decision-making rather than merely satisfying curiosity. Explanations should empower supervisors to intervene, adjust, or override AI recommendations based on contextual judgment.

4. **Interdisciplinary:** Drawing on insights from law, ethics, psychology, sociology, and domain-specific expertise. This expertise helps translate technical outputs into different sectoral contexts.

5. **Multi-faceted:** Addressing not just "how" but "why," "what if," and "what should be." Comprehensive interpretation requires understanding not just AI processes but their purposes, consequences, and normative implications.

Together, these XAI 2.0 principles recast explainability as a social and institutional process rather than a purely technical property of models. They provide a compass for AI supervisory authorities: features to be embedded in supervisory workflows, supported by professional structures that bridge technical and contextual expertise.

[29] DeepMind Safety Research, 'Negative Results for Sparse Autoencoders On Downstream Tasks and Deprioritising SAE Research' (Medium, 26 March 2025) <https://deepmindsafetyresearch.medium.com/negative-results-for-sparse-autoencoders-on-downstream-tasks-and-deprioritising-sae-research-6cadcfc125b9> accessed 12 November 2025.

Interpretative supervision is the institutional expression of this shift. XAI 2.0 articulates a more pragmatic notion of explainability, and interpretative supervision operationalises it within supervisory institutions. Its purpose is not to require that developers produce systems that are fully transparent or human-interpretable in advance, but to ensure that supervisors possess the institutional capacity to interpret AI behaviour in context, to determine whether deployed systems operate consistently with law, ethics, and democratic values. In this mode, AI supervision becomes an ongoing practice of contextual judgment, supported by technical expertise, stakeholder engagement, and institutional memory.

This broader vision parallels previous developments in other regulatory domains. Financial oversight has moved beyond static rulebooks to dynamic stress testing.[30] Public health has embraced rapid response capabilities for regulating novel technologies under uncertainty.[31] Across these fields, regulation has become more dynamic, outcome-focused, and interpretive, recognising that complex systems cannot be managed by rigid checklists and audits alone. AI supervision should build upon these models and capabilities.

The OBSERVE Framework: Institutionalising Interpretative Capacity for AI Supervision

Interpretative supervision defines what effective AI oversight must achieve: the institutional capacity to interpret, contextualise, and act upon AI behaviour under uncertainty. The OBSERVE framework sets out how this capacity can be built and sustained in practice. It translates the commitments of XAI 2.0 into organisational design: embedding interpretative functions within regulatory agencies through specialised units, technical infrastructures, and participatory networks. Each element of OBSERVE operationalises a facet of interpretative supervision: continuous observation, behavioural monitoring, stakeholder engagement, institutional learning, adaptive regulation, value integration, and early warning. Together, they form a coherent architecture for supervision that is dynamic, accountable, and aligned with democratic governance. This section outlines these components in turn, providing clarity on their structure, function, and the relationships that bind them into a coherent system of supervision.

[30] 'The Bank of England's Approach to Stress Testing the UK Banking System' (Bank of England 2015).
[31] 'Final Report on the Adaptive Pathways Pilot' (European Medicines Agency 2016).

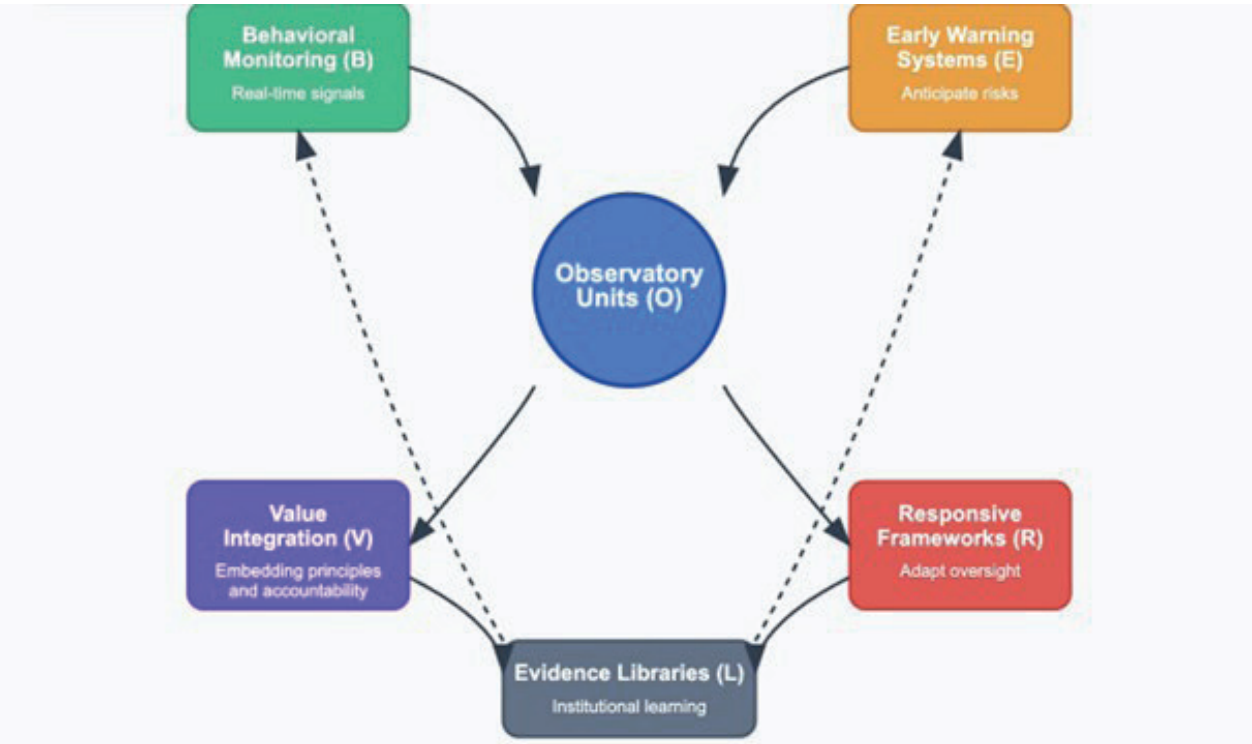

Fig. 1 Schematic representation of the OBSERVE framework.
Note: Continuous arrows show the primary information pathways, while the dashed ones represent the feedback loops.

O – Observatory Units

XAI 2.0 Feature: Context-based & Actionable

Observatory units are specialised structures within regulatory agencies that provide continuous institutional capacity to interpret AI system behaviour and translate it into enforceable regulatory action. They are specialised hubs embedded in regulatory agencies, staffed by hybrid teams of technical experts, domain specialists, and legal practitioners. Unlike compliance divisions that rely on periodic audits, observatories are designed for continuous situational awareness of AI deployments.

Their function is to integrate diverse data sources (e.g. operational benchmarks, adversarial stress tests, and user complaints) into a regulatory interpretation layer. This layer translates technical anomalies (e.g. robustness degradation, statistical parity shifts, anomaly scores) into legally enforceable judgments. A financial regulator's observatory, for instance, might detect non-linear drift in loan approvals correlated with macroeconomic stress, and determine whether this constitutes unlawful discrimination under fair lending law. A transport regulator could analyse near-miss clusters in autonomous vehicle telemetry and issue targeted safety directives before accidents occur.

Observatory units must have direct escalation powers into enforcement. They should not function as back-office analysts but as operational supervisors with the authority to trigger investigations, mandate remedial action, or suspend unsafe systems. In the EU, observatories align with the supervisory functions set out in the AI Act for market surveillance authorities (MSAs), meaning post-market monitoring (Art. 72), and incident notifications (Art. 73).[32]

B – Behavioral Monitoring

XAI 2.0 Feature: Falsifiable and Actionable

Behavioural monitoring is the function that ensures continuous oversight of AI systems by collecting and analysing data on their performance, risks, and anomalies during deployment. Behavioural monitoring addresses the mismatch between adaptive AI and static oversight.

Certification at a single point in time cannot guarantee ongoing AI safety.[33] Supervisors, therefore, require real-time supervisory technology (SUPTech) capable of monitoring AI systems as they evolve.[34] This entails establishing regulatory telemetry pipelines, where high-risk AI systems periodically or continuously report operational data to supervisors. Metrics should be carefully chosen to reflect compliance obligations: for example, fairness indices mapped to anti-discrimination law,[35] robustness thresholds linked to safety standards, anomaly detection for manipulation attempts, or systemic risk indicators in financial applications.

The challenge for this function lies in signal extraction as opposed to raw data collection. Monitoring systems must enrich anomalies with contextual interpretation, for instance, whether a rise in false positives is random noise, evidence of data drift, or a signal of adversarial attack. These systems must also be adaptive, allowing supervisors to add new monitored indicators dynamically as novel risks emerge, without overhauling the regulatory codebase. In practice, behavioural monitoring should function like the central banks' curation of real-time stress indicators: a constant stream of supervisory intelligence that enables early, proportionate intervention.[36]

[32] 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), ELI: http://data.europa.eu/eli/reg/2024/1689/oj.
[33] Pavel Dolin and others, 'Statistically Valid Post-Deployment Monitoring Should Be Standard for AI-Based Digital Health' (arXiv.org, 6 June 2025) <https://arxiv.org/abs/2506.05701v2> accessed 13 November 2025.
[34] Matt Grasser, 'Artificial Intelligence in Suptech: The Need for Public Sector Adoption and Adaptation' <https://www.mattgrasser.com/research/artificial-intelligence-in-suptech-the-need-for-public-sector-adoption-and-adaptation> accessed 13 November 2025.
[35] 'Algoprudence Repository' <https://algorithmaudit.eu/algoprudence/case-repository/> accessed 13 November 2025.
[36] Sebastian Doerr, Jose Maria Serena and Leonardo Gambacorta, 'Big Data and Machine Learning in Central Banks' (Bank for International Settlements).

To ensure that monitoring captures the most salient risks, behavioural pipelines should be structured into specialised clusters:

- **Fairness Monitoring Cluster**–continuously tracks distributional outcomes, disparate impacts, and protected-class fairness indices in decision-making systems.
- **Robustness & Reliability Cluster**–measures degradation under shifting data distributions, adversarial vulnerabilities, and resilience to environmental variability.
- **Manipulation & Integrity Cluster**–detects persuasive, deceptive, or adversarial system behaviours such as prompt injection, misinformation spread, or stealth manipulation.
- **Systemic & Market Risk Monitoring Cluster**–evaluates macro-level indicators such as correlated failures in financial markets, cascading disruptions in critical infrastructure, and monopolistic behaviour where dominant AI providers consolidate systemic power or distort competition.

S–Stakeholder Networks

XAI 2.0 Feature: Context-based and Interdisciplinary

Stakeholder networks are structured ecosystems that bring external expertise into regulatory processes to expand interpretative capacity. These networks should convene professionals across four key domains:

- *Domain experts* such as doctors, teachers, or financial analysts, who can articulate what AI outputs mean within specific sectoral standards and practices.
- *Technical AI experts,* who can formulate robust hypotheses on how and why systems make particular choices, highlighting failure modes and capability limits.
- *Social impact advisors,* including ethicists and social scientists, who can identify broader societal consequences such as discrimination, polarisation, or labour displacement.
- *Legal scholars,* who situate AI behaviours within constitutional and statutory frameworks, clarify whether outputs comply with legal obligations.

The same AI output can mean radically different things depending on context.[37] A medical AI's diagnostic recommendation, for instance, may appear to supervisors as a probability score, but to clinicians it carries implications for patient safety, treatment standards, and clinical ethics.[38] [39] Embedding domain expertise directly into supervision ensures that authorities are not forced to master every field AI touches, but can instead rely on professional networks to support them in interpreting behaviour through the lens of real-world impact.

For such networks to thrive, participation must be genuinely reciprocal. Experts should find tangible value in contributing, through access to anonymised case data, oversight findings, or controlled sandbox environments that enable empirical testing and shared learning. Their input must be recognised publicly, with contributions acknowledged in reports and policy guidance that reinforce the authority and visibility of expert judgment. Engagement should be flexible, supported by digital collaboration platforms and rotating working groups that allow participation without full-time commitment.

To remain credible, these networks must be insulated from regulatory capture. Rotational appointments, term limits, and conflict-of-interest disclosures would constitute helpful safeguards to protect the network's independence while maintaining continuity. A strong precedent exists in UNESCO's AI Ethics Experts Without Borders initiative[40], which unites specialists from over fifty countries to provide policy guidance and capacity building. Networks designed with similar openness and integrity allow supervisors to convene expertise rapidly, integrate diverse perspectives into their reasoning, and retain public confidence. In doing so, they ensure that interpretive supervision rests on a foundation of both technical competence and public accountability, a form of collective intelligence on par with the complexity of the systems it seeks to oversee.

E – Evidence Libraries

XAI 2.0 Feature: Falsifiable & Multi-faceted

Evidence libraries are repositories of regulatory knowledge that document past cases, technical failures, and supervisory responses to support institutional learning and consistency.[41] They should capture technical failure modes (e.g., reward hacking, adversarial vulnerability, drift-induced bias), interpretative methodologies used to assess them, stakeholder perspectives on their impacts, and consequent regulatory inter-

[37] Steven Alter, 'Understanding Artificial Intelligence in the Context of Usage: Contributions and Smartness of Algorithmic Capabilities in Work Systems' (2022) 67 International Journal of Information Management 102392.
[38] Karim Lekadir and others, 'FUTURE-AI: International Consensus Guideline for Trustworthy and Deployable Artificial Intelligence in Healthcare' (2025) 388 BMJ e081554.
[39] Jie Zhang and Zong-ming Zhang, 'Ethics and Governance of Trustworthy Medical Artificial Intelligence' (2023) 23 BMC Medical Informatics and Decision Making 1.
[40] 'AI Ethics Experts Without Borders | Global AI Ethics and Governance Observatory' <https://www.unesco.org/ethics-ai/en/ai-ethics-experts-without-borders> accessed 13 November 2025.
[41] E.g.: Tommy Shaffer Shane, 'AI Incident Reporting: Addressing a Gap in the UK's Regulation of AI' (Centre for Long-Term Resiliance) <https://www.longtermresilience.org/reports/ai-incident-reporting-addressing-a-gap-in-the-uks-regulation-of-ai/> accessed 13 November 2025.

ventions taken along with their outcomes. Over time, libraries enable supervisors to identify recurrent patterns, refine interpretative frameworks, and avoid repeating ineffective responses.

There is already a strong foundation in the literature for how such infrastructures could be designed. For example, the OECD's proposal for a common AI incident reporting framework sets out 29 standardised criteria that could underpin cross-border knowledge-sharing.[42]

Meanwhile, analysis of the AI Incident Database, which has catalogued over 900 cases of AI-related harms,[43] demonstrates both the feasibility and value of systematic incident collection, while also surfacing challenges such as taxonomic ambiguity and incomplete reporting.

The AI Act already provides a legal foundation for such efforts in the EU. Article 71 requires providers and public authorities to register information on high-risk AI deployments in an EU database. With systematic expansion by Market Surveillance Authorities and curation by the AI Office, this infrastructure could evolve into a comprehensive evidence library: documenting not only system registrations but also incidents, supervisory responses, and their outcomes. Properly resourced, it would serve as the backbone of regulatory learning across the Union, and in time could be federated into a global incident-sharing ecosystem coordinated by conveners such as UNESCO or the OECD.

Global precedents from other industries also show how such infrastructures could be scaled across different jurisdictions. In aviation, the ICAO ADREP/ECCAIRS systems[44] pool global accident and incident data, allowing supervisors to spot systemic risks and prevent recurrences. In medicine, the WHO's VigiBase aggregates over 30 million adverse drug reaction reports from 170+ countries[45], enabling early detection of safety signals and coordinated regulatory action.

R – Responsive Frameworks

XAI 2.0 Feature: Actionable and Falsifiable

Responsive frameworks are adaptive regulatory mechanisms that adjust oversight intensity and methods in line with evolving risks and evidence. They replace rigid input rules with adaptive, outcome-oriented oversight. Standards should focus on key problem areas such as fairness and robustness, while allowing relative flexibility in how they are achieved. Frameworks must be regularly updated in light of new evidence and literature. Key instruments include:

[42] OECD, 'Towards a Common Reporting Framework for AI Incidents' (34th edn, 2025) OECD Artificial Intelligence Papers <https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html> accessed 13 November 2025.
[43] Kevin Paeth and others, 'Lessons for Editors of AI Incidents from the AI Incident Database' (arXiv.org, 24 September 2024) <https://arxiv.org/abs/2409.16425v1> accessed 13 November 2025.
[44] 'Home Page - ICAO Data+' <https://dataplus.icao.int/> accessed 13 November 2025.
[45] 'VigiBase Data Access | UMC' <https://who-umc.org/vigibase-data-access/> accessed 13 November 2025

- **Regulatory sandboxes**–controlled environments where new AI applications can be trialled under enhanced supervision. These create a safe space for innovation while allowing supervisors to gather empirical evidence on risks.

- **Dynamic risk assessments**–oversight intensity that scales with evolving deployment risks, informed by monitoring, incident reporting, and contextual factors.

- **Stress tests**–simulation of extreme but plausible conditions (e.g. large-scale data poisoning, cascading model failures, coordinated cyberattacks) that expose vulnerabilities and test resilience.

Different jurisdictions are already experimenting with responsive mechanisms in practice. In the European Union, the AI Act Articles 57 and 58 establish regulatory sandboxes and real-world testing environments as formal tools of adaptive oversight. In parallel, the European Central Bank's stress tests provide a mature model for dynamic supervision, scaling regulatory interventions in proportion to systemic risk, while the European Union Agency for Cybersecurity (ENISA) conducts cross-border cyber exercises to test resilience beyond finance.[46] In the United States, the NIST Management Framework provides a voluntary but widely adopted evidence-based structure for managing AI risks across the lifecycle.[47] Singapore's AI Verify offers a government-backed testing and verification framework that allows companies to demonstrate compliance with AI principles in a structured and supervised setting.[48]

At the international level, responsive oversight has already proven its value in financial governance. The Financial Stability Board (FSB) pioneered systemic stress-testing frameworks in response to the 2008 financial crisis, a model that has since become a permanent feature of global financial regulation.[49] These methods have since been extended to other systemic risk domains, including cybersecurity and climate resilience, illustrating how adaptive oversight can evolve into institutionalised global practice.

Supervisory authorities in AI governance can draw directly from these precedents. Stress-test targets should be identified through a tiered, evidence-based approach, prioritising high-risk or systemically significant AI systems, meaning those with extensive deployment, critical societal impact, or recurrent compliance alerts. Observatory units and behavioural monitoring pipelines can supply the intelligence to identify such systems, ensuring that tests focus on genuine points of systemic vulnerability rather than arbitrary samples. This

[46] 'Cross-Sector Exercise Requirements' (European Union Agency for Cybersecurity 2022) <https://data.europa.eu/doi/10.2824/941158> accessed 13 November 2025.
[47] AI Risk Management Framework' (NIST, 12 July 2021) <https://www.nist.gov/itl/ai-risk-management-framework> accessed 13 November 2025.
[48] 'What Is AI Verify' (AI Verify Foundation) <https://staging.aiverifyfoundation.sg/what-is-ai-verify/> accessed 13 November 2025.
[49] 'History of the FSB' (Financial Stability Board) <https://www.fsb.org/about/history-of-the-fsb/> accessed 13 November 2025.

mirrors the logic of financial supervision, where oversight is directed according to exposure, interdependence, and risk drift, enabling supervisors to allocate resources efficiently and maximise public value.[50]

V – Value Integration

XAI 2.0 Feature: Interdisciplinary and Multi-faceted

Values integration is the practice of embedding principles and public accountability into the supervision of AI systems. It ensures that AI governance reflects societal values and ethical commitments essential to public trust. In the AI domain, two global normative frameworks are especially significant. The OECD AI Principles, adopted by 46 countries in 2019 (and updated in 2024), established widely endorsed commitments to human-centred values, fairness, transparency, robustness, and accountability.[51] They have since informed the G20 AI Principles and provided a foundation for national AI strategies. Complementing this, the UNESCO Recommendation on the Ethics of AI, adopted unanimously by 193 member states in 2021, sets out the first global standard on AI ethics. It emphasises human rights, dignity, environmental sustainability, and inclusivity, and calls for regulatory mechanisms to embed these values in practice.[52]

Effective values integration requires mechanisms that balance technical assurance with democratic legitimacy. Ethical-impact assessments should, for instance, rigorously examine how AI systems affect autonomy, dignity, and trust, ensuring that mitigation measures do not compromise fundamental rights. Citizen-engagement and redress mechanisms should grant individuals meaningful avenues to challenge algorithmic decisions that affect them, reinforcing accountability and procedural fairness. Crucially, interpretative frameworks for AI governance should remain open to public deliberation and periodic review, allowing supervision to evolve alongside societal and technological change. In this way, values integration anchors AI supervision into a broader democratic and constitutional function, aligning governance of emerging technologies with enduring commitments to human rights, justice, and public accountability.

E – Early Warning Systems

XAI 2.0 Feature: Multi-faceted & Falsifiable

Early warning systems are anticipatory mechanisms that detect emerging risks and capabilities before they manifest in deployment. Unlike observatories, which focus on interpreting anomalies in deployed systems, or behavioural monitoring, which detects AI performance in real time, early warning systems look outward and forward, scanning the horizon for transformative capabilities and systemic threats that may destabilise entire sectors if left unaddressed.

[50] 'Core Principles for Effective Banking Supervision' (Bank for International Settlements 2024).
[51] 'AI Principles' (OECD) <https://www.oecd.org/en/topics/ai-principles.html> accessed 13 November 2025.
[52] 'Recommendation on the Ethics of Artificial Intelligence' (UNESCO 2021) <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence> accessed 13 November 2025.

Building effective early warning systems requires horizon-scanning pipelines that continuously analyse signals across research publications, patent filings, open-source repositories, venture capital flows, and industry disclosures. By triangulating these sources, supervisors can detect breakthroughs in areas such as biological simulation, synthetic media, or autonomous reasoning architectures that may radically alter the risk landscape. Detection must then be translated into scenario planning and structured war-gaming. Supervisors should run foresight exercises to model how frontier capabilities might interact with critical domains under stress.

Concrete AI foresight efforts already exist. The UK Government Office for Science's AI 2030 Scenarios Report[53] provides decision-useful narratives and critical uncertainties for policymakers. The series of International Scientific Reports on the Safety of Advanced AI, synthesises expert judgment on frontier risks and priority evaluation areas. CSET's Emerging Technology Observatory (ETO)[54] offers a practical example of how signals from papers, patents, and code can be transformed into actionable early-warning indicators.

To build on these efforts, supervisors should embed dedicated foresight units within agencies, expand scanning to include economic and organisational signals such as investment flows and compute use, and link AI foresight to existing systems in health, finance, and cybersecurity. Internationally, a global foresight observatory under neutral conveners (e.g. UNESCO, OECD, and G7) could consolidate national pipelines and enable cross-border intelligence sharing. Structured public–private partnerships, including safe-harbour regimes for voluntary risk disclosures, would further strengthen the collective radar. In this way, early warning systems become the forward flank of interpretative supervision: scanning for signals, stress-testing consequences, and enabling collective anticipation.

Implementation and Limitations of the Framework

The OBSERVE framework can be implemented incrementally. Agencies with limited capacity might begin by appointing an AI oversight lead, maintaining a simple case library, and joining international peer networks. As they develop, they can establish small observatory units, adopt basic monitoring tools, and convene stakeholder groups. More well-resourced supervisors can invest in full-scale real-time supervision, institutionalise stress testing, and lead international coordination of early warning systems. In this way, interpretative supervision becomes accessible to supervisors at all levels of capacity while building toward a shared global architecture of AI governance.

[53] 'AI 2030 Scenarios Report HTML (Annex C) - GOV.UK' (UK Department for Science, Innovation & Technology 2024) <https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/ai-2030-scenarios-report-html-annex-c> accessed 13 November 2025.
[54] 'Emerging Technology Observatory' <https://eto.tech/> accessed 13 November 2025.

While the OBSERVE framework provides a structured pathway for scaling interpretative supervision, its application will inevitably face constraints. In many settings, resource scarcity, data asymmetries, and limited access to proprietary systems may restrict the depth of interpretative analysis and the speed of supervisory response. Moreover, interpretative supervision operates within ongoing power and information imbalances between authorities and developers, where key technical insights remain commercially sensitive or strategically withheld. Implementing OBSERVE will therefore require not only organisational reform but also sustained negotiation over access, accountability, and the sharing of supervisory intelligence. Finally, the framework should not be understood as a fixed blueprint but as an adaptive scaffold evolving with context, incorporating local governance traditions, and remaining open to revision as both AI technologies and societal expectations continue to change.

Conclusion

The challenge of governing artificial intelligence is at once technical, societal and institutional. AI's opacity, adaptivity, and systemic reach demand forms of supervision that move beyond static audits or mechanistic transparency. Supervising authorities must be equipped to make sense of AI in context to interpret its outputs in ways that are legally enforceable, socially grounded, and democratically accountable.

The OBSERVE framework offers a pathway for building this capacity. Its seven elements – Observatory Units, Behavioural Monitoring, Stakeholder Networks, Evidence Libraries, Responsive Frameworks, Values Integration, and Early Warning Systems – translate the features of XAI 2.0 into concrete institutional designs. Together, they form an architecture that is context-based, falsifiable, actionable, interdisciplinary, and multifaceted. In practice, this means supervisory authorities can detect harms earlier, respond proportionately, and align AI systems with public values, even as technologies evolve in unpredictable directions.

Crucially, OBSERVE is scalable and adaptive. Low-capacity agencies can begin with simple case libraries and participation in international peer networks, mid-tier supervisors can establish observatory units and stakeholder ecosystems, and well-resourced jurisdictions can institutionalise real-time monitoring, stress testing, and foresight coordination. Through this progressive layering of capability, interpretative supervision evolves from local experimentation to global practice, fostering a shared architecture of accountability and trust in the governance of AI.

References

Alter S, 'Understanding Artificial Intelligence in the Context of Usage: Contributions and Smartness of Algorithmic Capabilities in Work Systems' (2022) 67 International Journal of Information Management 102392

Batool A, Zowghi D and Bano M, 'AI Governance: A Systematic Literature Review' (2025) 5 AI and Ethics 3265
'Core Principles for Effective Banking Supervision' (Bank for International Settlements 2024)

'Cross-Sector Exercise Requirements' (European Union Agency for Cybersecurity 2022) <https://data.europa.eu/doi/10.2824/941158> accessed 13 November 2025

Doerr S, Serena JM and Gambacorta L, 'Big Data and Machine Learning in Central Banks' (Bank for International Settlements)

Dolin P and others, 'Statistically Valid Post-Deployment Monitoring Should Be Standard for AI-Based Digital Health' (arXiv.org, 6 June 2025) <https://arxiv.org/abs/2506.05701v2> accessed 13 November 2025

Elgin CY and Elgin C, 'Ethical Implications of AI-Driven Clinical Decision Support Systems on Healthcare Resource Allocation: A Qualitative Study of Healthcare Professionals' Perspectives' (2024) 25 BMC Medical Ethics 148

'Final Report on the Adaptive Pathways Pilot' (European Medicines Agency 2016)

Floridi L (ed), Ethics, Governance, and Policies in Artificial Intelligence, vol 144 (Springer International Publishing 2021) <https://link.springer.com/10.1007/978-3-030-81907-1> accessed 12 November 2025

Gabriel I and Ghazavi V, 'The Challenge of Value Alignment: From Fairer Algorithms to AI Safety' (arXiv, 19 January 2021) <http://arxiv.org/abs/2101.06060> accessed 12 November 2025

Goos M and Savona M, 'The Governance of Artificial Intelligence: Harnessing Opportunities and Mitigating Challenges' (2024) 53 Research Policy 104928

Hyzen A and others, 'Epistemic Welfare and Algorithmic Recommender Systems: Overcoming the Epistemic Crisis in the Digitalized Public Sphere' [2025] Communication Theory qtaf018

'International Scientific Report on the Safety of Advanced AI: Interim Report' (Department for Science, Innovation and Technology and AI Safety Institute 2024) <https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai> accessed 13 November 2025

Janssen M, 'Responsible Governance of Generative AI: Conceptualizing GenAI as Complex Adaptive Systems' (2025) 44 Policy and Society 38

Judge B, Nitzberg M and Russell S, 'When Code Isn't Law: Rethinking Regulation for Artificial Intelligence' (2025) 44 Policy and Society 85
Kolt N, Shur-Ofry M and Cohen R, 'Lessons from Complex Systems Science for AI Governance' (2025) 6 Patterns 101341

Lekadir K and others, 'FUTURE-AI: International Consensus Guideline for Trustworthy and Deployable Artificial Intelligence in Healthcare' (2025) 388 BMJ e081554

Liza Mark and Tianyun (Joyce) Ji, 'China Publishes the AI Security Governance Framework' (Haynes and Boone 2024)

Longo L and others, 'Explainable Artificial Intelligence (XAI) 2.0: A Manifesto of Open Challenges and Interdisciplinary Research Directions' (2024) 106 Information Fusion 102301

OECD, 'Towards a Common Reporting Framework for AI Incidents' (34th edn, 2025) OECD Artificial Intelligence Papers <https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html> accessed 13 November 2025

Okolo CT, 'Governance of AI-Based Algorithms', Handbook of Human-Centered Artificial Intelligence (Springer, Singapore 2025) <https://link.springer.com/rwe/10.1007/978-981-97-8440-0_84-1> accessed 12 November 2025

Paeth K and others, 'Lessons for Editors of AI Incidents from the AI Incident Database' (arXiv.org, 24 September 2024) <https://arxiv.org/abs/2409.16425v1> accessed 13 November 2025

Pedreschi D and others, 'Human-AI Coevolution' (arXiv, 6 May 2024) <http://arxiv.org/abs/2306.13723> accessed 12 November 2025

'Recommendation on the Ethics of Artificial Intelligence' (UNESCO 2021) <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence> accessed 13 November 2025

Seger E, 'In Defence of Principlism in AI Ethics and Governance' (2022) 35 Philosophy & Technology 45

Shane TS, 'AI Incident Reporting: Addressing a Gap in the UK's Regulation of AI' (Centre for Long-Term Resiliance) <https://www.longtermresilience.org/reports/ai-incident-reporting-addressing-a-gap-in-the-uks-regulation-of-ai/> accessed 13 November 2025
Taeihagh A, 'Governance of Generative AI' (2025) 44 Policy and Society 1

'The Bank of England's Approach to Stress Testing the UK Banking System' (Bank of England 2015)

Wilhelmina Afua Addy and others, 'AI in Credit Scoring: A Comprehensive Review of Models and Predictive Analytics' (2024) 18 Global Journal of Engineering and Technology Advances 118

World Health Organization, Systems Thinking: For Health Systems Strengthening (World Health Organization 2009)

Zeng J, Artificial Intelligence with Chinese Characteristics: National Strategy, Security and Authoritarian Governance (Springer 2022) <https://link.springer.com/10.1007/978-981-19-0722-7> accessed 12 November 2025

Zhang J and Zhang Z, 'Ethics and Governance of Trustworthy Medical Artificial Intelligence' (2023) 23 BMC Medical Informatics and Decision Making 1

SECTION II.

AI SANDBOXES AND OTHER EVIDENCE-BASED TESTING FOR COMPETENT AUTHORITIES

Authorities have highlighted that AI supervision must be informed by empirical understanding, by seeing how AI systems behave under real conditions, how risks manifest, and how regulatory requirements translate into practice. Sandboxes and other evidence-based testing environments have therefore emerged as central instruments for building institutional learning and regulatory certainty.

This theme occupied a prominent place at the Expert Roundtable, reflecting the widespread demand among authorities for concrete guidance on how to design, manage, and benefit from sandbox initiatives. The discussions revealed that sandboxes are no longer perceived merely as experimental spaces for innovators but as structured learning environments for supervisors themselves. Participating in or operating a sandbox helps authorities and developers alike clarify how the AI Act applies in practice, how definitions are interpreted, standards are applied, and obligations are operationalised.

Speakers pointed out that "sandboxing" is not a uniform practice but a family of approaches serving different purposes. Some sandboxes focus on technical validation, testing robustness, accuracy, or safety, while others emphasise regulatory interpretation and process design. Many combine both dimensions.[1] Examples from Brazil,[2] Singapore,[3] and Thailand[4] illustrated this diversity, showing how sandboxes can range from large-scale financial infrastructures to targeted pilots in healthcare or education. What unites them is the shift from abstract compliance to evidence-driven supervision, where supervisors engage directly with real systems and real data.

Several cross-cutting lessons emerged. Timing and institutional culture are critical: successful sandboxes depend on readiness within authorities to experiment, share information, and tolerate a degree of uncertainty. Capacity-building remains indispensable, particularly for public officials trained in risk-averse administrative traditions. Moreover, inclusive participation – of civil-society actors, technical experts, and end-users – enhances legitimacy and broadens the evidentiary base for regulation.

The following shed light on these different aspects from different perspectives: Brazil's experience with regulatory sandboxes as a form of institutional experimentalism, the European EUSAiR project's roadmap for AI Act implementation through coordinated sandbox networks, and the Datasphere Initiative's analysis of sandboxes as instruments for building public trust Together, these analyses provide a concrete picture of what evidence-based supervision can look like in practice.

[1] Sophie Tomlinson, 'Sandboxes and AI Innovation in Europe' (The Datasphere Initiative, 1 May 2025) <https://www.thedatasphere.org/news/sandboxes-and-ai-innovation-in-europe/> accessed 13 November 2025.
[2] Rodrigo de Oliveira Andrade, 'Boost for Startups: Government Passes Bill to Encourage Investment in Small Innovative Companies in Brazil' <https://revistapesquisa.fapesp.br/en/boost-for-startups/> accessed 13 November 2025.
[3] 'Generative AI Evaluation Sandbox' (Infocomm Media Development Authority) <https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/generative-ai-evaluation-sandbox> accessed 13 November 2025.
[4] 'Insights from Practice: Building an AI Governance Clinic in Thailand | Global AI Ethics and Governance Observatory' <https://www.unesco.org/ethics-ai/en/articles/insights-practice-building-ai-governance-clinic-thailand> accessed 13 November 2025.

EU REGULATORY SANDBOXES FOR AI

Authors: Fabio Seferi, Antonino Rotolo, Marco Billi; EUSAiR Project

Abstract

This paper analyses the introduction and operationalisation of AI regulatory sandboxes under Article 57 of the EU Artificial Intelligence Act, mandating Member States to establish controlled environments that foster innovation while ensuring compliance with fundamental rights, health, and safety standards. Drawing on empirical findings and multi-stakeholder engagement within the EUSAiR project, this contribution analyses key operational, legal, and institutional design parameters essential for effective sandbox implementation. The paper points to key persisting challenges, including the lack of unified compliance tools compatible with AI Act requirements, disparities in national oversight capacities, and limited synergies with existing horizontal regulatory regimes, such as data protection and product safety laws. In this respect, the paper stresses the need for the adoption of a modular sandbox architecture, differentiated between advisory, technical, and testing-intensive phases, and its integration into the overall EU AI innovation ecosystem, encompassing the European Digital Innovation Hubs, Testing and Experimentation Facilities, AI Factories, and High-Performance Computing initiatives. Furthermore, it advocates for legal interoperability protocols, mitigating fragmentation and allowing for mutual recognition of sandbox activities across jurisdictions. By framing the concept of sandboxes not as one-shot tests but rather as iterative, dynamic regulatory processes, the paper advances practical proposals positioning regulatory sandboxes both as innovation accelerators and normative laboratories underpinning the future evolution of EU digital regulation in the AI domain.

[1] https://eusair-project.eu. For more information, send an email to info@eusair-project.eu.

A New Frontier in AI Governance

The Concept of AI Regulatory Sandboxes

The legislative framework put in place by the Artificial Intelligence Act of the European Union introduces AI regulatory sandboxes as one of the fundamental tools to achieve innovation while maintaining compliance with legal requirements. Article 57(1) requires each Member State to establish at least one AI regulatory sandbox at the national level by 2 August 2026. AI regulatory sandboxes are defined, under article 3(55) of the AI Act, as "a controlled framework set up by a competent authority which offers providers or prospective providers of AI systems the possibility to develop, train, validate and test, where appropriate in real-world conditions, an innovative AI system, pursuant to a sandbox plan for a limited time under regulatory supervision".

The configuration of the sandboxes may be physical venues, a fully digital platform, or hybrid models able to host a wide range of AI applications, from pure software solutions to AI embedded in physical products. This flexibility in space and technology allows for tailored supervision that is adapted to the specificities of different AI innovations.

AI regulatory sandboxes have several objectives: first, they offer a controlled environment to experiment within, hence decreasing legal uncertainty for innovators and easing entry barriers, especially for startups and SMEs; second, they serve as a means for competent authorities to build a sophisticated understanding of the capabilities and risks of AI technologies, thereby informing regulatory learning and increasing policy responsiveness; and third, they hasten the innovation-to-market cycle since they promote proactive problem-solving and early identification of regulatory compliance challenges.[2]

The EUSAiR project of the European Union represents a concrete answer to the mandate of the AI Act, implemented by close cooperation with the European AI Office. EUSAiR develops standardised yet adaptable frameworks for sandbox implementation across Member States, focusing on capacity building, cross-jurisdictional cooperation, and the integration of sandbox activities within the broader AI innovation ecosystem. Based on its empirical research and multi-stakeholder engagement, the project has advanced key theoretical and operational insights into how to catalyse the evolution of EU digital regulation in this key domain.

The EU AI Innovation Ecosystem: A Landscape of Opportunities

The EU AI Innovation Ecosystem represents a vast and interconnected network of actors providing a diverse range of services. These services span digital maturity assessments, specialised training programs,

[2] For more details, we refer the reader to (Bagni and Seferi, 2025).

comprehensive AI Act guidance, robust infrastructure, advanced computing resources, cutting-edge testbeds, and essential AI components. Effectively leveraging this rich ecosystem is crucial for the successful operation of AIRS. Such an ecosystem is made up of several key initiatives.

A first key initiative is represented by the European Digital Innovation Hubs (EDIHs).[3] EDIHs serve as regional and thematic entry points for digital transformation, offering tailored services to SMEs, public authorities, and other organisations. They combine technological expertise with business support to accelerate digital uptake, especially in under-digitised regions and sectors.

Second, another important actor is the Euro High Performance Computing Joint Undertaking (EuroHPC JU).[4] EuroHPC JU is a public-private initiative to position the EU as a global leader in supercomputing and quantum computing. It seeks to provide world-class infrastructure and foster innovation in computation-intensive sectors and applications. Coupled to this, AI Factories serve as centralised infrastructures for AI model development, offering advanced computing resources and a suite of services that facilitate the training, testing, deployment, and ongoing maintenance of general-purpose and application-specific AI systems.[5]

Third, testing and experimentation facilities (TEFs) are also key.[6] TEFs are large-scale environments dedicated to testing, validating, and demonstrating advanced AI systems, both software and hardware, in real or close-to-real conditions. They are central to de-risking AI deployment and aligning with regulatory frameworks such as the AI Act.

Fourth, the AI-on-Demand Platform (AIoDP) may offer a wide range of services, with its most valuable feature being the platform itself.[7] It serves as a collaborative space for the AI community, enabling engagement with peers and experts, sharing opportunities, applications, and knowledge, and accessing AI-related assets and tools.

Fifth, Data Spaces are designed as sector-specific data ecosystems that enable secure, sovereign, and interoperable data sharing among stakeholders.[8] They establish a governance and trust framework for data exchange in compliance with EU regulations, particularly in relation to data protection, competition law, and ethical use.

Finally, there are also upcoming initiatives under the AI Continent Plan.[9] The AI Continent Plan outlines the establishment of five Gigafactories, each designed to house over a million advanced processors

[3] More information available at: https://european-digital-innovation-hubs.ec.europa.eu/it/home.
[4] More information available at: https://www.eurohpc-ju.europa.eu/index_en.
[5] More information available at: https://european-digital-innovation-hubs.ec.europa.eu/it/home.
[6] More information available at: https://digital-strategy.ec.europa.eu/en/faqs/testing-and-experimentation-facilities-tefs-questions-and-answers.
[7] More information available at: https://www.aiodp.eu/.
[8] More information available at: https://digital-strategy.ec.europa.eu/en/policies/data-spaces.
[9] More information available at: https://commission.europa.eu/topics/eu-competitiveness/ai-continent_en.

dedicated to the development and training of sophisticated AI models. Moreover, the plan anticipates the operational launch of 15 AI Factories by 2026, each with a specific sectoral focus and equipped with a network of advanced antennas.

Considering all of the above, this paper aims to provide a general overview of the main challenges and opportunities in the establishment of AI regulatory sandboxes, synthesising the findings of EUSAiR for the benefit of supervisory authorities and policymakers.[10] The following sections build a coherent analytical pathway: Section 2 identifies the operational, regulatory, and methodological elements underlying the development of sandboxes; Section 3 explores synergies between sandboxes and the EU's AI innovation ecosystem, thus clarifying strategic alignments that can be conducive to effective supervision and support for innovation.

Navigating the Current Challenges in Establishing AI Regulatory Sandboxes

Operational and Regulatory Challenges in AI Regulatory Sandbox Implementation

Setting up AI regulatory sandboxes within the European Union requires an intricate balancing of challenges pertaining to design issues, models of operation, stakeholder engagement, and integrating them within wider ecosystems. The operational framework of sandboxes needs to be flexible to accommodate the full AI development life cycle, from early advisory support to advanced conformity assessments, and different stages would require different regulatory interventions. Consequently, sandboxes will not only have to provide initial compliance checking mechanisms but also iterative reassessments and adaptations of the AI systems in line with technological readiness and considerations of business maturity.

Yet, one of the fundamental challenges lies in the availability and interoperability of AI compliance tools. Current toolsets are, in many aspects, not well-adapted to the subtleties of the AI Act and may lead to inconsistent interpretation and enforcement in different jurisdictions. Closing these gaps requires coordinated development of complete and modular compliance frameworks, with the ability to support novice and advanced AI providers alike inside sandboxes.

Another important aspect is to ensure wide accessibility and to incentivise participation. The European AI Office considers it particularly important to promote, and when necessary, provide free of charge,

[10] We refer the interested reader to them for a more extended discussion. Section 1 is based on EUSAiR Roadmap (EUSAiR 2025a) and offers an overview of some major issues and challenges. Section 2 briefly recalls contributions from (EUSAiR 2025b) and (EUSAiR 2025c); : it identifies the services as 'building blocks' of AI regulatory sandboxes and discusses types of testing and regulatory challenges. Section 3 elaborates in brief on some more detailed analysis offered in a second publicly accessible report (EUSAiR 2025b) on synergies with the innovation ecosystem. Some conclusions end the paper.

access to the sandbox for SMEs and startup businesses in order to boost innovation in a truly inclusive and vibrant manner. Yet, the trade-off between incentivization, administrative burden, and supervisory capacity remains, which requires further action with a view to resource allocation, procedural transparency, and ongoing stakeholder outreach.

Ecosystem Integration, Governance, and Methodological Foundations for Effective Sandbox Supervision

Regulatory sandboxes do not operate alone; they interact with the overall EU constellation of innovation infrastructures in Europe, such as EDIHs, TEFs, AI Factories, and Data Spaces. It is important to achieve maximum synergies among them for better sharing of infrastructures, knowledge, and technical expertise. This also calls for legal interoperability protocols and operational standards to avoid fragmentation and provide for mutual recognition of sandbox activities across Member States.

Governance structures must clearly outline the roles and responsibilities of competent authorities, ensuring that oversight of health and safety and fundamental rights is well-coordinated with sectoral and horizontal regulatory regimes, including data protection, intellectual property, and consumer safety laws. The sustainability of sandbox operations depends on sufficient human and financial resources being invested in training programs, capacity building, and continuous professional development of regulatory personnel.[11]

Regulatory sandboxes do not operate alone; they interact with the overall EU constellation of innovation infrastructures in Europe, such as EDIHs, TEFs, AI Factories, and Data Spaces. It is important to achieve maximum synergies among them for better sharing of infrastructures, knowledge, and technical expertise. This also calls for legal interoperability protocols and operational standards to avoid fragmentation and provide for mutual recognition of sandbox activities across Member States.
Governance structures must clearly outline the roles and responsibilities of competent authorities, ensuring that oversight of health and safety and fundamental rights is well-coordinated with sectoral and horizontal regulatory regimes, including data protection, intellectual property, and consumer safety laws. The sustainability of sandbox operations depends on sufficient human and financial resources being invested in training programs, capacity building, and continuous professional development of regulatory personnel.

[11] For more details, we refer the reader to (Novelli et al., 2025).

The EU AI Ecosystem and AI Regulatory Sandboxes: Potential Synergies

Aligning AIRS with Ecosystem Offerings: Mapping the Synergies

The expansive EU AI Innovation Ecosystem can be strategically leveraged for the effective execution of several key AIRS building blocks. We have identified significant complementarity in the potential services offered by AI Regulatory Sandboxes and the diverse range of actors within the EU AI Innovation Ecosystem. This complementarity can play a critical role in establishing a robust and effective EU-wide network of AI Regulatory Sandboxes and fostering impactful cross-regional AIRS.

AIRS Building Block	Actor	Actor's Related Service
Guidance on regulatory expectations (Scope, Definitions, Risk Classification, etc.)	EDIH (Sector-agnostic)	• AI Maturity Assessment (In the upcoming EDIH round) • Training and upskilling for AI providers
Guidance on AI Act Compliance	EDIH (Sector-agnostic)	• AI Act Helpdesk (In the upcoming EDIH round) • Training and upskilling
	TEFs (Sectoral)	• Provide guidance on AI Act alignment
Guidance on Health, Safety and Fundamental Rights risks identification, testing, and mitigation	TEFs (Sectoral)	• Integration and validation of AI systems • Physical and virtual testbeds • Technical compliance testing
	AI Factories (Cross-sectoral)	• AI technical experimentation environments offering controlled space for development, training, and evaluation of AI systems • Technical documentation and developer support • AI testing, validation, and compliance • assistance

Table 1: AIRS Synergies with the EU AI Ecosystem's Services

Continuation of Table 1: AIRS Synergies with the EU AI Ecosystem's Services

AIRS Building Block	Actor	Actor's Related Service
Development, Training, Testing	Data Spaces (Sectoral)	• Data governance frameworks • Interoperability standards • Facilitation of data trading and sharing
	AI on Demand Platform	• Experimentation Services: exploring, • testing, and creating with shared AI tools on the platform • Enabling collaboration with other teams • Accessing existing datasets, AI tools and services, use cases, scientific publications, funding opportunities, training resources, and upcoming AI-related events.
	AI Factories (Cross-sectoral)	• AI-optimised compute and storage infrastructure • Training and fine-tuning of models • Technical experimentation sandboxes offering controlled environments for development, training, and evaluation of AI systems • Technical documentation and developer support • AI testing, validation, and compliance assistance • In-house management of software environments and data resources
	Euro HPC (Cross-sectoral)	• Access to top-tier supercomputing facilities • Support for R&D in HPC and quantum computing
	TEFs (Sectoral)	• Physical and virtual testbeds • Compliance testing and AI Act alignment
Monitoring and acting on significant risks to Health, Safety and Fundamental Rights	TEFs (Sectoral)	• Progress tracking via structured models (e.g., "graduation" schemes)
Facilitating development of tools, benchmarks for accuracy, robustness, cybersecurity, etc.	AI Factories (Cross-sectoral)	• Technical experimentation sandboxes offering controlled environments for development, training, and evaluation of AI systems • Hosting benchmark testing

Continuation of Table 1: AIRS Synergies with the EU AI Ecosystem's Services

AIRS Building Block	Actor	Actor's Related Service
Facilitating development of measures to mitigate risk, Health, Safety and Fundamental Rights	TEFs (Sectoral)	• Integration and validation of AI systems • Physical and virtual testbeds • Compliance testing and AI Act alignment
	AI Factories (Cross-sectoral)	• Technical experimentation sandboxes offering controlled environments for development, training, and evaluation of AI systems • Documentation and developer support • AI testing, validation, and compliance assistance
Compliance with other EU & national regulations and involvement of DPA and other NCAs	-	-
Testing in real-world conditions	TEFs (Sectoral)	• Demonstration and real-world experimentation
Conformity Assessment Preparation	TEFs (Sectoral)	• Compliance testing and AI Act alignment • Progress tracking via structured models (e.g., "graduation" schemes)

The table illustrates that the majority of AIRS building blocks (see Section 2.1) align with both current and planned services offered within the EU AI Ecosystem by various actors. This underscores the critical importance of cohesive coordination to ensure streamlined and effective implementation across the entire European Union.

Some AIRS services could potentially be provided by multiple actors, which could lead to overlaps. For example, "Development, Training, Testing" is a service that includes diverse components which need to be considered separately, and different actors can offer different components of the service. The AI on Demand Platform enables participants to access datasets, experiments, and try pit AI tools. Data Spaces also offer sectoral datasets, which should be aligned with the participant's sector of operations. EuroHPC and AI Factories can both provide infrastructure for AI development, training, and testing. AI Factories, however, offer a broader set of services which enable alignment with ethical

and legal frameworks. Furthermore, TEFs can also provide testing services, but with a deeper AI Act alignment and real-world testing conditions, albeit only in four sectors. In such cases, NCAs are advised to carefully align the specific needs of the AIRS participant with the sectoral scope, technological focus, geographic location, and level of regulatory guidance required by the targeted actor.

A notable gap exists, as the building block focusing on "Compliance with other Union and national regulations as well as engagement with Data Protection Authorities (DPAs) and other NCAs" is currently not addressed by any existing stakeholder. This presents a challenge that NCAs must address by developing this capacity internally.

AI Ecosystem's Positioning Along the AIRS Operational Pipeline: A Streamlined Journey

Drawing from the synergies identified between the AIRS building blocks and the services available within the EU AI Innovation Ecosystem, several feasible scenarios emerge for implementing AIRS across Member States. These institutional and infrastructural synergies are underscored by stakeholder feedback, which is the main concern in operationalising these connections, particularly the degree of coordination among ecosystem actors such as EDIHs, TEFs, the AI-on-Demand Platform, Data Spaces, EuroHPC, and AI Factories. To align with the AI Office's Guidance Note on the common sandbox process[12], we propose a structured framework, the AIRS Pipeline, that follows the same sequential steps. It is recommended that AIRS be underpinned by a well-defined framework that enables a streamlined and successful client journey, particularly for SMEs:

- Pre-participation: Initial assessment and preparation.
- Application & Selection: Formal submission and review process.
- Preparation: Detailed planning and resource allocation.
- Participation: Active engagement in sandbox activities.
- Evaluation & Exit: Assessment of results and conclusions.
- Post-participation: Follow-up and continued support.
- Reporting & Monitoring: Ongoing oversight and analysis.

[12] Commission Staff Working Document, 'Regulatory learning in the EU Guidance on regulatory sandboxes, testbeds, and living labs in the EU, with a focus section on energy', SWD(2023) 277/2 final.

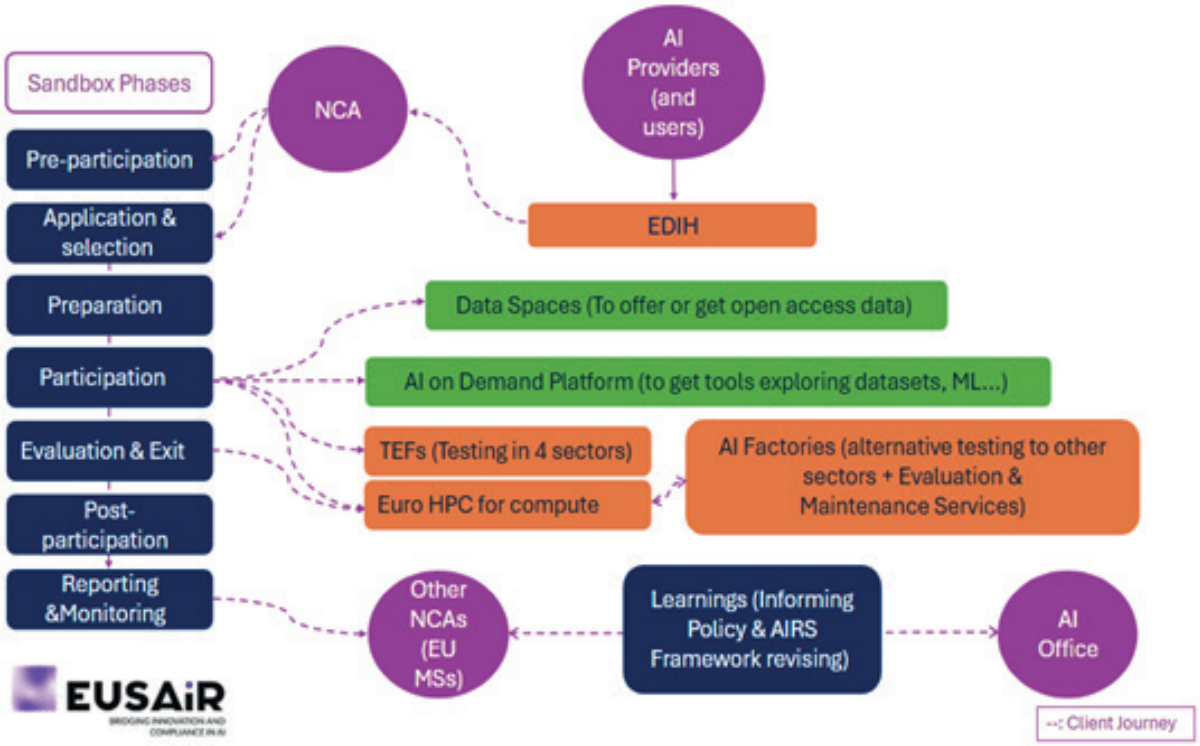

Figure 1: The EU AI Innovation Ecosystem Pipeline for AIRS1

We envision EDIHs as the first point of contact in the AIRS Pipeline, ensuring that the needs of SMEs remain central. EDIHs are well-positioned to conduct initial assessments of technical, legal, and market readiness, advise on funding opportunities, and guide innovators towards suitable next steps – whether that means AIRS participation or preparatory support.

For example, an AI provider at an early ideation stage might first benefit from EDIH training services, financing advice, or access to "test before invest" environments before entering an AIRS. In this sense, EDIHs play a triaging role, ensuring that only AI systems with sufficient technical maturity and a relevant regulatory challenge proceed into the AIRS process. When appropriate, EDIHs can also liaise directly with NCAs to communicate preliminary assessments and readiness findings

The AI-on-Demand platform can serve as a valuable resource, granting access to datasets, experimentation opportunities, and AI tools.[13] TEFs can provide comprehensive support for AI testing, including real-world conditions, while Euro HPC infrastructures and AI Factories can offer advanced computing solutions.

[13] The AI-on-Demand Platform is designed, for researchers and innovators, to offer an extensive repository of over 500,000 AI assets-including datasets, tools, and software. For industry, including SMEs, startups, and the public sector, the platform delivers market-ready AI solutions and robust support across the entire innovation lifecycle, from low-code development tools and HPC training resources to secure European hosting environments and a dedicated AI marketplace. Find more information at https://www.eit.europa.eu/news-events/news/europes-ai-demand-platform-powering-research-and-innovation

Which Level of Technological Maturity of AI for Regulatory Sandboxes: TRL Scale

The Technology Readiness Level (TRL) scale is a common method providing a standardised metric for assessing the maturity of AI technologies.[14]

- **TRL 1-2: Ideation and Conceptualisation:** Technologies are in the early stages of conceptual development.
- **TRL 3-7: Prototyping and Validation:** Development transitions from concept to functional prototypes.
- **TRL 8-9: Real-world Validation and Late-Stage Compliance:** Solutions are deployed and evaluated in operational environments.

It is important to acknowledge that the decision of when to engage developers depends on their current TRL maturity. It is important to underscore that flexibility in accepting different TRL levels might be advisable for encouraging participation from companies with diverse innovation processes, as some may require support earlier in their development journey, while others will engage at advanced stages.

AIRS Building Block	Minimum TRL
Guidance on regulatory expectations (Scope, Definitions, Risk Classification, etc.)	Min TRL 2-3
Guidance on AI Act Compliance	Min TRL 2-3
Guidance on H, S, FR risks identification, testing, mitigation and their effectiveness	Min TRL 2-3
Development, Training, Testing	Min TRL 3
Monitoring and acting on significant risks to H, S, and FR arising during Sandbox operations	Min TRL 3
Facilitating development of tools, benchmarks for accuracy, robustness, cybersecurity, etc.	Min TRL 4
Facilitating the development of measures to mitigate risk to H, S, FR	Min TRL 4
Compliance with other Union & national regulations and involvement of DPA and other NCAs	Min TRL 2-3
Testing in real-world conditions	Min TRL 4
Conformity Assessment Preparation	Min TRL 5-6

Table 2: Alignment of AIRS Services with TRL Thresholds

[14] Version as described by the International Organization for Standardization (ISO) with the publication of the ISO 16290:2013 standard. Find more information at Mihály Héder, From NASA to EU: the evolution of the TRL scale in Public Sector Innovation, The Innovation Journal: The Public Sector Innovation Journal, Volume 22(2), 2017, article 3.

The use of the TRL scale is not enough; it is crucial that business maturity, market potential and funding should be considered during the AIRS criteria when selecting an actor.

Stage-Gate Model for AIRS Execution: Integrating Business and Technical Readiness

Adopting a Stage-Gate model that incorporates both TRLs and business readiness is recommended for guiding the execution of AIRS. A Stage-Gate model is a structured project management approach that divides the development and evaluation of an AI system into distinct stages separated by gates. Each stage focuses on specific assessment tasks, such as technical validation, risk classification, and compliance evaluation, while each gate serves as a decision point where authorities determine whether the system can advance to the next stage. In the context of AIRS, this model ensures that AI systems are systematically assessed for maturity, safety, and regulatory compliance, beginning with eligibility and risk classification (Gates 0–1) and progressing through detailed technical and ethical evaluations (Gates 2–6) before market entry. This model can assist authorities in evaluating AI solutions, determining the necessary services, and deciding whether the solution is ready for market entry.

Business maturity, market potential, and funding are closely linked stages in an AI company's lifecycle. Business maturity describes a company's established and stable state, typically indicated by a reliable customer base and steady revenue, which, for innovative AI systems, could be decisive. This is also connected to the company's capability, including expertise, necessary resources (technology) and financing (funding), and the overall earning logic, feasibility, soundness and legality of the business model. Market potential concerns the growth opportunities available in the market, indicating how much a company can expand its sales or reach. Funding, finally, provides the financial resources needed for the company to achieve its objectives, whether that is expanding operations, investing in new AI products, or entering new markets. For example, an AI system with a high TRL might not succeed if the business lacks the market potential or required funding.

To address this two-fold analysis, the solution is to consider both TRLs and business readiness indicators, evaluating both technical and business maturity elements, such as market potential and funding. For example, a start-up with innovative AI technology but lacking market insight will not fare well unless it's also assessed for business readiness. This dual framework would facilitate the alignment of sandbox activities with the building blocks that could be provided by key AI ecosystem stakeholders across the EU.[15]

[15] For more details, we refer the reader to (Heikkilä 2025) and (EUSAiR 2025b).

Conclusion

The successful implementation of AI regulatory sandboxes across the EU will require a sustained and well-coordinated effort to address the identified challenges and leverage the potential synergies between the sandboxes and the broader EU AI ecosystem. The EUSAiR project has been working on delivering a valuable roadmap for establishing effective sandboxes, offering guidance on regulatory expectations, compliance support, and facilitating development and testing. From our experience so far, it is clear that NCAs can create robust and context-sensitive AIRS that promote AI innovation, improve legal certainty, and ensure compliance with the AI Act. Continued efforts to address the remaining gaps in the ecosystem, such as the lack of comprehensive compliance tools and the uneven geographic distribution of key infrastructure, will be essential for realising the full potential of AI regulatory sandboxes in driving responsible AI innovation and growth across the European Union. By embracing a collaborative and forward-thinking approach, the EU can position itself as a global leader in AI governance and innovation.

References

Bagni, F. and Seferi, F., eds. 2025. “Regulatory sandboxes for AI and Cybersecurity. Questions and answers for stakeholders.” CINI’s Cybersecurity National Lab. https://cybersecnatlab.it/white-paper-on-regulatory-sandboxes/.

EUSAiR. 2025a. “EU Regulatory Sandboxes for AI: EUSAiR Roadmap.” April. https://eusair-project.eu/library/eusair-roadmap/.

—. 2025b. “AI Ecosystem and AI Regulatory Sandboxes.” July. https://eusair-project.eu/library/ai-ecosystem-and-ai-regulatory-sandboxes/.

—. 2025c. “Pilot Guide: Why and how to join the EUSAiR Pilots.” July. https://eusair-project.eu.

Heikkilä, J., Lagstedt, A., Heikura, S., Kaakkurivaara, E., Dardykina, A., & Teilhard de Chardin, A. 2025. “From Simple Sandbox Process to Regulatory Sandbox Framework: Serving the Dual Objectives of AI Regulation.” 38th Bled eConference: Empowering Transformation: Shaping Digital Futures for All: Conference Proceedings. Univerzitetna založba Univerze v Mariboru. 643-660.

Novelli, C., Hacker, P., McDougall, S., Morley, J., Rotolo, A. and Floridi, L. 2025. “Getting Regulatory Sandboxes Right: Design and Governance Under the AI Act.” https://ssrn.com/abstract=5332161 or http://dx.doi.org/10.2139/ssrn.5332161.

INSTITUTIONAL EXPERIMENTALISM FOR AI SUPERVISION: FOSTERING PUBLIC-SECTOR INNOVATION THROUGH SANDBOXES

Author: Lucas Costa Dos Anjos; Agência Nacional de Proteção de Dados (ANPD - Brazilian Data Protection Authority)

Abstract

This contribution analyses Brazil's experience with regulatory sandboxes as a lens for understanding how public-sector institutions can build the capacity required for effective AI supervision. It argues that sandboxes should not be viewed as deregulatory shortcuts or simple innovation accelerators, but as catalysts for institutional experimentalism, a process through which supervisory authorities develop new reflexes, governance tools, and organisational cultures suited to emerging technologies. Drawing on lessons from the Central Bank of Brazil and the Data Protection Authority, the paper highlights the practical barriers civil servants face when engaging in experimentation, including legal uncertainty, institutional risk aversion, and procedural inertia. It proposes a framework for embedding experimental governance within public institutions, emphasising principled innovation, cultural change, and structured learning processes. In doing so, the contribution positions sandboxes as instruments not only for testing novel AI systems but for strengthening state capacity and shaping a more adaptive, responsive, and accountable AI governance ecosystem.

Introduction

Artificial intelligence has thrust supervisory authorities into uncharted territory. The question is no longer whether to regulate emerging technologies, but how to do so while fostering responsible innovation and safeguarding public protection. Regulatory sandboxes have gained considerable traction as a policy tool, promising controlled environments where innovative solutions can be tested within relaxed regulatory constraints.[1]

The Brazilian experiences with regulatory sandboxes (both with the Central Bank and the Data Protection Authority) offer a sobering case study. What emerges from examining their journey is not always a straightforward success story, but rather a complex picture of institutional learning curves and regulatory innovations. Public servants, it turns out, do not naturally gravitate towards experimentation. The very idea of regulatory sandboxes can trigger deep-seated anxieties about legal liability, institutional reputation, and departure from time-tested procedures.[2]

This analysis proposes an "institutional experimentalism," a framework that acknowledges organisational realities while building capacity for principled innovation. Rather than sidestepping bureaucratic resistance through deregulatory shortcuts, this approach invests in the cultural and procedural changes necessary for sustainable experimental governance.[3] The stakes are particularly high for AI supervision, where technological complexity, market concentration, and societal implications demand active state engagement rather than passive accommodation of private sector experimentation. The central argument here challenges the prevalent notion that regulation hampers innovation. Ideally, supervisory authorities might transform themselves from reluctant gatekeepers into active shapers of technological development, but only if they commit to genuine institutional learning and capacity-building, rather than superficial regulatory flexibility.

[1] Oskar Josef Gstrein and Anne Beaulieu, ‚Regulatory sandboxes in the AI Act: reconciling innovation and safety?' (2023) 49 Computer Law & Security Review 359. See also: James Vieira, B., Campo, P. M. A., Alcântara, R. R. B. de & Melo, C. A. V. de, ‚O sandbox regulatório como instrumento de incentivo à inovação no Brasil: os casos do Banco Central do Brasil, da Comissão de Valores Mobiliários e da Superintendência de Seguros Privado' (2024) 153(1) Revista do TCU 343.
[2] World Bank, Global experiences from regulatory sandboxes, Finance, Competitiveness & Innovation Global Practice (Fintech Note No. 8, International Bank for Reconstruction and Development 2020).
[3] David A Wolfe, ‚Experimental Governance: Conceptual approaches and practical cases' (OECD Local Economic and Employment Development Papers, OECD Publishing 2020).

The panacea problem of when sandboxes become silver bullets

The word "panacea" carries weight in Brazilian policy discussions about regulatory sandboxes. Derived from the Greek goddess of healing, it captures how these frameworks have been marketed as solutions to virtually every challenge facing regulator-regulated relationships. This framing sets unrealistic expectations and obscures the substantial institutional work required to make experimental governance effective.[4]

Brazil's Central Bank discovered this firsthand. This Central Bank's sandbox initiative followed international precedents, particularly the UK's pioneering Financial Conduct Authority sandbox in 2016. The timing was significant: Brazil implemented its first sandbox during a period of rapid fintech expansion and increasing pressure to modernise financial regulation, positioning itself as Latin America's most ambitious experiment in financial regulatory innovation. Unlike the more established European frameworks, Brazil's approach unfolded in an emerging market context characterised by high market concentration, limited regulatory resources, and pressing developmental imperatives that made the stakes of regulatory experimentation considerably higher. Of 52 initial project submissions in their first sandbox cycle, only seven received authorisations to proceed, which is a substantial filtering that may reveal the gap between policy rhetoric and operational reality.[5] This also reflects not merely selective evaluation criteria, but deeper institutional challenges in translating experimental governance concepts into workable frameworks.

Brazil's approach to AI governance through regulatory sandboxes presents another variation on this theme of ambitious policy rhetoric confronting institutional reality. The Agência Nacional de Proteção de Dados (ANPD) launched its AI regulatory sandbox programme before Brazil even enacted comprehensive AI legislation, creating what might appear as a proactive regulatory response to emerging technology challenges.[6] The ANPD's approach demonstrates a more structured response to experimental governance than initial Central Bank efforts. Rather than treating algorithmic transparency as a peripheral concern, the Authority established transparency as the primary object of its sandbox programme, requiring participants to develop "techniques and technologies that promote algorithmic transparency" as an eliminatory criterion for participation.[7]

[4] Quatrochi, G. et al., ‚Banks 4.0 in Brazil: possibilities to ensure fintechs financing role through its market positioning' (2023) 13 (3) Innovation and development 561.
[5] Banco Central do Brasil, Relatório de Gestão - Sandbox Regulatório: 1° Ciclo 2022 (Banco Central do Brasil 2022).
[6] ANPD, ‚ANPD's Call for Contributions to the regulatory sandbox for artificial intelligence and data protection in Brazil is now open' (3 October 2023).
[7] ANPD, ‚Edital n° 2/2025: Participação em Piloto de Ambiente Regulatório Experimental em Inteligência Artificial e Proteção de Dados' (Diário Oficial da União, 27 June 2025) art 3°, art 19(IV).

This focus represents a shift from deregulatory accommodation toward active regulatory shaping of AI development priorities. The programme's design incorporates mandatory capacity-building through a structured "levelling phase" comprising theoretical and practical training sessions to ensure participants understand regulatory methodology before testing commences. Such institutional investment in participant preparation suggests recognition that effective experimental governance requires systematic knowledge transfer.

The ANPD's sandbox focuses on machine learning-driven technologies and generative AI systems, operating as what officials describe as "a controlled environment to test technologies associated with artificial intelligence developed by participants" while implementing "good practices to ensure compliance with personal data protection rules and principles."[8] This framing reveals the familiar tension between innovation promotion and regulatory compliance that characterises most sandbox initiatives. The institutional architecture supporting Brazil's AI sandbox demonstrates its ambition, since it was developed in collaboration with the Development Bank of Latin America and the Caribbean.[9] The programme drew on regional methodologies and experience, suggesting recognition that sandbox design requires substantial institutional learning rather than simple policy transplantation.[10] A public consultation process attracted 71 contributions from various stakeholders, including private entities, civil society organisations, academia, and public sector bodies, and the projected timeline of 18 to 24 months from proposal to final report publication is an indication of awareness that experimental governance requires sustained institutional commitment rather than quick regulatory fixes.[11]

International recognition followed, with the World Economic Forum highlighting Brazil's initiative alongside sandbox programmes in Singapore, the United States, the United Arab Emirates, the United Kingdom, the European Union, and Mauritius.[12] The programme's design reflects familiar aspirations to balance technological innovation with fundamental rights protection, particularly privacy and data protection. Like Brazil's Central Bank experience, the ANPD's AI sandbox confronts the challenge of translating experimental governance concepts into workable institutional frameworks while managing the gap between policy ambitions and operational realities.

The institutional reality proves far messier than policy documents may suggest. Civil servants trained in traditional regulatory approaches, with their emphasis on legal certainty, risk mitigation, and pro-

[8] ANPD, ‚ANPD's Call for Contributions to the regulatory sandbox for artificial intelligence and data protection in Brazil is now open' (3 October 2023).
[9] CAF, the Development Bank of Latin America and the Caribbean, has played a strategic role in promoting innovative regulatory frameworks for artificial intelligence across the region. The institution has supported several Latin American countries in developing AI regulatory sandboxes, including pioneering initiatives in Brazil, Colombia and Chile. Armando Guío, ‚Regulatory Sandbox for Artificial Intelligence in Chile: Discussion Document' (Development Bank of Latin America and the Caribbean/CAF, August 2021).
[10] Data Protection & Privacy 2024 - Brazil' (Chambers and Partners 2024).
[11] ANPD, ‚ANPD prorroga prazo de inscrições para o sandbox regulatório de inteligência artificial e proteção de dados' (Autoridade Nacional de Proteção de Dados, 7 August 2025).
[12] ANPD, ‚ANPD's Regulatory Sandbox Featured at report of the World Economic Forum' (26 January 2024).

cedural compliance, often view experimental initiatives as fundamentally at odds with their professional responsibilities.[13] This creates a peculiar dynamic in which the very people charged with implementing sandbox policies may harbour serious reservations about their appropriateness or even their feasibility. Questions about legal liability, debates over regulatory competence boundaries, and scepticism about whether sandbox mechanisms address the regulatory challenges they purport to solve are all on the table.[14] These concerns reflect genuine institutional anxieties rather than mere bureaucratic stubbornness (though anyone who has worked in government will recognise there's usually some of both). They signal deeper tensions between experimental governance requirements and deeply established regulatory cultures.[15]

The panacea problem also extends beyond individual project selection to broader questions about regulatory modernisation. When sandboxes are positioned as comprehensive solutions, they risk becoming substitutes for more fundamental institutional reforms. This proves particularly problematic in emerging markets, where regulatory agencies often face capacity constraints that cannot be addressed through experimental flexibility alone.[16] Rather than building lasting capabilities for innovation governance, sandbox initiatives may create temporary exceptions that fail to generate systemic regulatory adaptation. Moreover, this cure-all perception can lead authorities to over-rely on private sector initiatives while diminishing their own role in shaping technological development. When regulatory agencies primarily function as facilitators of private sector innovation rather than active participants in technological governance, they may also inadvertently reinforce existing power asymmetries.[17] This dynamic proves especially concerning in sectors characterised by significant technological complexity and rapid change, like artificial intelligence, where maintaining independent regulatory expertise becomes of the utmost importance for effective oversight.

[13] Allen, H.J., ‚Regulatory sandboxes' (2019) 87 George Washington Law Review.
[14] Comitê Estratégico de Gestão do Sandbox BC (CESB), Ata 1ª Reunião extraordinária do Comitê Estratégico de Gestão do Sandbox Regulatório (26 March 2021).
[15] Lisa Yue, Babak Asgari and Olga Hawn, ‚Regulation and Innovation Revisited: How Restrictive Environments Can Promote Destabilizing New Technologies' (2024) 35(4) Organization Science 1449.
[16] Crampes, C. & Estache, A., ‚Efficiency vs. distributional concerns in regulatory sandboxes' (2025) Journal of economic policy reform 1.
[17] Mazzucato, M. and Collington, D., The Big Con: How the Consulting Industry Weakens Our Businesses, Infantilizes Our Governments and Warps Our Economies (Allen Lane 2023).

Bureaucratic resistance: the human factor in regulatory innovation

Understanding why civil servants resist experimental governance requires examining the institutional incentives and professional cultures that shape bureaucratic behaviour. The Brazilian experience may illuminate several key sources of resistance that supervisory authorities must address systematically, rather than simply overcome or circumvent. Legal interpretation emerges as a primary battleground for experimental governance. Civil servants receive extensive training in prioritising legal certainty and compliance with established frameworks. When experimental initiatives require departures from standard legal interpretations, bureaucrats naturally default to conservative positions that minimise perceived legal exposure. Brazil's Central Bank Legal Office analysis of sandbox applications demonstrated this tendency, with legal opinions consistently emphasising regulatory competence boundaries and discretionary limits.[18]

This legal conservatism reflects rational responses to asymmetric institutional incentives. Civil servants face potential personal and professional liability for regulatory decisions that subsequently prove problematic, creating powerful incentives to avoid experimental approaches that deviate from established precedents.[19] The career consequences of regulatory failures typically outweigh any rewards for successful innovation, systematically biasing bureaucratic decision-making against experimental initiatives.[20] Risk assessment within bureaucratic contexts tends to emphasise potential negative outcomes while discounting potential benefits, particularly when dealing with unfamiliar technologies or regulatory approaches. Such a dynamic is usually marked by discussions about potential market distortions, consumer protection concerns and systemic risks, with limited corresponding analysis of innovation benefits or experimental safeguards.[21] This risk-averse mindset reflects professional training and institutional cultures that often reward caution over creativity.

Procedural adaptation presents another dimension of bureaucratic resistance. Established regulatory procedures embody decades of institutional learning and professional practices. Experimental initiatives requiring procedural modifications may be perceived as threats to institutional competence and professional identity. Civil servants may resist experimental approaches not because they oppose innovation in principle, but because they view established procedures as superior mechanisms for achieving regulatory objectives. The resistance extends to institutional identity itself. Regulatory agencies develop organisational cultures emphasising institutional stability, predictability, and systematic application of regulatory requirements.[22]

[18] CESB, Ata 2ª Reunião ordinária do Comitê Estratégico de Gestão do Sandbox Regulatório (16 July 2021).
[19] William Occasio, Jeffrey Loewenstein and Arjun Nigam, ‚Institutional Logics: Motivating Action and Overcoming Resistance to Change' (2023) 19(6) Management and Organization Review 1171y.
[20] European Securities and Markets Authority and European Insurance and Occupational Pensions Authority, FinTech: Regulatory Sandboxes and Innovation Hubs (JC 2018 74, 2018).
[21] CESB, Ata 4ª Reunião ordinária do Comitê Estratégico de Gestão do Sandbox Regulatório (17 September 2021).
[22] Rainer Kattel, Wolfgang Drechsler and Erkki Karo, How to Make an Entrepreneurial State: Why Innovation Needs Bureaucracy (Yale University Press 2022).

Experimental governance frameworks prioritising flexibility, adaptation, agile culture and iterative learning may conflict with these institutional identities, creating cognitive dissonance for civil servants who identify strongly with more traditional regulatory approaches.[23] On one hand, organisational structures combine these individual-level sources of resistance. On the other hand, supervisory authorities typically organise around functional specialisations that may not align well with experimental initiatives' cross-cutting nature. Meanwhile, sandbox programmes require coordination across multiple departments, integration of diverse expertise, and adaptation of existing workflows, all of which can create organisational friction and resistance. Yet, this resistance should not be dismissed as mere obstruction. Civil servants' concerns about legal liability, procedural integrity, and institutional mission reflect legitimate institutional values that must be preserved even as agencies adapt to experimental governance requirements. The challenge lies in designing institutional frameworks that address these concerns while enabling principled experimentation.

Learning opportunities for regulators within a sandbox experience

Rather than viewing sandboxes merely as temporary accommodations for private sector innovation, forward-thinking regulators can use these experiences to build institutional capabilities, develop new expertise, and refine their approach to emerging technologies. The most immediate learning opportunity involves developing a technical understanding of emerging technologies. Regulatory staff are usually required to grapple with technical concepts, new business models, and risk profiles that extend beyond traditional regulated services and products. This is compatible with broader regulatory trends, including the emergence of AI literacy obligations in various jurisdictions as part of comprehensive AI governance frameworks. Supervisory authorities ought to develop competencies in understanding AI systems and their implications. This technical learning proves particularly valuable because it occurs within regulatory contexts rather than abstract academic settings. When civil servants evaluate blockchain-based certificate trading platforms or review algorithmic credit decisions, for example, they develop a real and practical understanding of how these technologies function in actual market conditions. Sandboxes thus serve as practical training grounds where regulatory staff can develop the AI literacy and technical fluency increasingly required by modern regulatory frameworks. This contextual knowledge becomes invaluable for future regulatory decisions, enabling more informed assessments of similar innovations and more sophisticated regulatory and enforcement frameworks.

[23] Kattel, R. et al., How to make an entrepreneurial state: why innovation needs bureaucracy (Yale University Press 2022). See also: David A Wolfe, ‚Experimental Governance: Conceptual approaches and practical cases' (OECD Local Economic and Employment Development Papers, OECD Publishing 2020).

Furthermore, sandbox experiences provide regulatory authorities with opportunities to refine their risk assessment capabilities. Traditional regulatory approaches often rely on standardised risk categories and historical precedents that may not capture the risk profiles of innovative business models. Through sandbox experimentation, regulators can observe how new technologies and business models perform under controlled conditions, developing a more nuanced understanding of their risk characteristics. Initial risk assessments based on theoretical concerns might be refined through actual operational experience, leading to a more sophisticated understanding of which risks materialise and which prove less significant than anticipated throughout the sandbox experience. This empirical approach to risk assessment enables regulators to move beyond precautionary assumptions towards more evidence-based regulation.

Also, sandboxes force regulatory institutions to examine and adapt their own internal processes. Regulators may evolve their evaluation procedures, monitoring mechanisms, and decision-making processes throughout the sandbox implementation. These procedural innovations could prove more lasting than individual project outcomes, creating institutional capabilities that extend beyond specific experimental initiatives. The learning extends to coordination mechanisms across different regulatory units. Sandbox programmes typically require input from legal, technical, policy, and supervisory departments, forcing institutions to develop better coordination mechanisms and cross-functional working relationships. Though challenging and difficult to operationalise, these organisational improvements often outlast specific sandbox cycles, creating more agile and responsive regulatory institutions in the future.

In addition to that, regulatory sandboxes necessitate different types of stakeholder engagement than traditional regulatory processes. Regulators must work closely with innovators, monitor ongoing operations, and facilitate learning exchanges between different participants. This intensive engagement develops regulatory staff's capabilities for collaborative governance approaches that prove valuable across broader regulatory functions. Initial interactions are usually focused primarily on compliance and risk mitigation, but may gradually expand to include discussions about market development, innovation trajectories, and policy implications. This more sophisticated stakeholder engagement capability enhances regulatory institutions' ability to anticipate and shape technological developments.

Perhaps most importantly, sandbox experiences provide regulatory authorities with empirical foundations for broader policy development. Rather than developing regulations based entirely on theoretical considerations or international best practices, regulators can draw on direct experience with how specific technologies and business models function within their jurisdictions. Insights about technology applications, digital innovations, and alternative business models generated through sandbox experimentation can subsequently influence broader approaches to technology regulation.

Finally, beyond specific technical or procedural learning, sandbox experiences can also contribute to broader institutional culture change. Successful sandbox implementations demonstrate that experimental approaches can coexist with regulatory rigour, gradually building institutional comfort with innovation-oriented regulatory approaches. This culture change often occurs gradually, through the accumulation of positive experiences, rather than by means of a single dramatic institutional transformation. As civil servants gain confidence with experimental approaches and observe successful outcomes, institutional resistance to innovation-oriented regulatory frameworks diminishes. This cultural evolution proves crucial for sustaining experimental governance beyond specific sandbox initiatives.

From traditional innovation to AI supervision: institutional lessons for institutional experimentalism

AI governance presents both opportunities and challenges for institutional experimentalism. While technological contexts may differ substantially from traditional regulation, the institutional dynamics of existing sandbox experiences offer valuable guidance for AI supervisory authorities seeking to implement effective experimental governance frameworks. AI supervision presents even greater technical complexity, requiring regulatory staff to understand machine learning algorithms, data processing practices, algorithmic decision-making systems, and their societal implications. Regulatory institutions must then invest substantially in technical education and cross-disciplinary collaboration to develop adequate oversight capabilities.[24] The learning curve for AI supervision proves steeper than traditional regulation because AI systems continue evolving through ongoing training, deployment in new contexts, and interaction with other systems. Unlike static products and services that can be evaluated once and approved, AI systems require ongoing monitoring and adaptive oversight approaches that match their dynamic nature.[25] AI markets exhibit even greater concentration tendencies than traditionally regulated services, raising concerns about how experimental governance frameworks might reinforce existing power asymmetries. Regulatory relief can thus disproportionately benefit well-capitalised incumbents who possess the legal expertise and resources necessary to navigate experimental frameworks.[26] Therefore, AI supervisory authorities must carefully consider how experimental governance approaches might exacerbate market concentration in sectors already dominated by large technology firms. The technical barriers to AI development, combined with network effects and data advantages, create market dynamics that may be more resistant to competition-enhancing regulatory interventions than traditional markets.

[24] Just, S. N. et al., ‚Open for business: the discursive diffusion of regulatory sandboxes for fintech innovation' (2024) 17 (3) Journal of cultural economy 360.
[25] OECD, ‚Regulatory sandboxes in artificial intelligence' (2023) OECD Digital Economy Papers, No. 356.
[26] OECD, Regulatory Sandboxes for Privacy -- Analytical Report (Business at OECD (BIAC) 2020).

The public interest implications of AI systems extend far beyond market efficiency concerns that traditionally dominate regulated fields.[27] AI systems affect privacy, discrimination, democratic participation, and fundamental rights in ways that require broader consideration of societal values and democratic accountability. This expansion of regulatory scope demands institutional capabilities that extend beyond technical expertise to include social impact assessment and democratic engagement. Moreover, AI governance presents greater needs for international regulatory coordination than domestic regulation. AI systems often operate across national boundaries, and the global nature of major AI developers creates very complex jurisdictional challenges, which corroborates the need for AI supervisory authorities to develop even more sophisticated coordination mechanisms that extend internationally.[28]

Institutional experimentalism offers AI supervisory authorities a systematic approach to managing the tension between innovation promotion and effective oversight. The idea is to address the organisational challenges that emerge when traditional regulatory agencies attempt experimental governance while maintaining regulatory rigour and public interest orientation, beginning with the establishment of clear legal authorities and procedural frameworks that provide civil servants with confidence about experimental governance's legitimacy and boundaries. AI supervisory authorities benefit tremendously from explicit statutory authorities for experimental activities, defined boundaries for regulatory discretion, and clear protections against personal liability for good-faith experimental decisions.[29]

These legal frameworks must address the unique characteristics of AI systems, including their evolving nature, cross-sectoral applications, and international scope. Unlike other fields that can be regulated within established sectoral boundaries, AI systems often span multiple regulatory domains, requiring legal frameworks that enable coordination across agencies while maintaining clear lines of authority and accountability. Supervisory authorities must establish dedicated experimental governance units with clear mandates, sufficient resources, and explicit authority to coordinate across organisational boundaries. These coordination mechanisms must balance experimental flexibility with accountability requirements. While experimental governance requires some departure from standard regulatory procedures, it cannot operate without systematic monitoring, evaluation, and public reporting.[30] AI supervisory authorities must then establish transparent evaluation criteria, regular reporting requirements, and clear mechanisms for stakeholder input and oversight.

[27] Washington, P. B. & Lee, E., ‚Nexus between regulatory sandbox and performance of digital banks: A study on UK digital banks' (2022) 15 (12) Journal of risk and financial management 1.
[28] Anjos, L., ‚A bird's-eye view of the Paris AI Action Summit: Regulation, power, and alternatives' Tech Global Institute (25 February 2025). See also: Markos Chatzipanagiotou and Ioannis Koulierakis, ‚AI Governance in a Complex and Rapidly Changing Regulatory Landscape: A Global Perspective' (2024) 11 Humanities and Social Sciences Communications 1134.
[29] Lima, C. M. de & Pasqualetto, A., ‚Installation of Regulatory Sandbox Environment from the Perspective of the Brazilian Charter for Smart Cities' (2023) 11 (83) Revista Nacional de Gerenciamento de Cidades.
[30] David A Wolfe, ‚Experimental Governance: Conceptual approaches and practical cases' (OECD Local Economic and Employment Development Papers, OECD Publishing 2020).

Traditional regulatory training also proves insufficient for AI governance challenges, requiring systematic investment in technical education, cross-disciplinary collaboration skills, and adaptive management capabilities. AI supervisory authorities must establish ongoing professional development programmes that combine technical AI education with regulatory innovation methodologies. This capacity building extends beyond individual skills to encompass institutional capabilities. The Brazilian experience demonstrates the importance of deliberate scale limitations for experimental governance programmes. ANPD restricted participation in its sandbox to "up to three participants over twenty months," suggesting conscious recognition that experimental governance effectiveness depends on institutional capacity to provide intensive oversight rather than extensive programme reach.[31] This approach contrasts with sandbox initiatives that prioritise broad industry accommodation over deep institutional learning. Regulatory agencies require technical infrastructure for monitoring AI systems, analytical capabilities for processing complex data about algorithmic behaviour, and institutional processes for coordinating across different areas of expertise. Building these capabilities requires sustained investment and strategic planning rather than ad hoc responses to immediate challenges.[32]

Additionally, institutional experimentalism requires sophisticated risk management approaches that distinguish between acceptable experimental risks and unacceptable public interest harms. Unlike traditional risk management that seeks to minimise all potential negative outcomes, experimental risk management must carefully calibrate acceptable uncertainty levels while maintaining protection against serious harms.[33] Therefore, AI supervisory authorities must develop risk assessment frameworks specifically designed for experimental governance that include clear escalation procedures, intervention triggers, and termination criteria. These frameworks must address both technical risks related to AI system performance and broader societal risks related to privacy, discrimination, and democratic participation.

Experimental governance requires ongoing dialogue with multiple stakeholder communities, including industry participants, civil society organisations, academic researchers, and affected communities.[34] These engagement processes must inform experimental design, provide ongoing feedback during implementation, and contribute to evaluation and learning processes. For AI governance, stakeholder engagement becomes particularly complex because AI systems affect diverse communities in different ways, often with asymmetric power relationships and varying levels of technical understanding. Supervisory authorities must then develop engagement mechanisms that enable meaningful participation from affected communities while maintaining technical rigour and regulatory independence.

[31] ANPD, ‚Edital n° 2/2025: Participação em Piloto de Ambiente Regulatório Experimental em Inteligência Artificial e Proteção de Dados' (Diário Oficial da União, 27 June 2025).
[32] Cheng, S.-Y. & Hou, H., ‚Innovation, financial development, and growth: evidences from industrial and emerging countries' (2022) 55 (3) Economic change and restructuring 1629.
[33] World Bank, Global experiences from regulatory sandboxes, Finance, Competitiveness & Innovation Global Practice (Fintech Note No. 8, International Bank for Reconstruction and Development 2020).
[34] David A Wolfe, ‚Experimental Governance: Conceptual approaches and practical cases' (OECD Local Economic and Employment Development Papers, OECD Publishing 2020).

Managing the state-market dynamic in AI supervision

The relationship between state authorities and market participants in AI experimental governance presents great complexity, given AI systems' broad societal implications and the concentration of AI development capabilities within a small number of global technology firms. In this scenario, regulatory agencies risk becoming overly dependent on private sector technical assessments when they lack independent expertise.[35] This dependency proves particularly problematic in AI governance, where the complexity of systems and the rapid pace of development can overwhelm regulatory capabilities.

The ideal is that AI supervisory authorities systematically invest in building internal technical capabilities that enable independent evaluation of AI systems and their societal implications. This investment extends beyond hiring technically trained staff to include ongoing education, technical infrastructure, and institutional processes that maintain a cutting-edge understanding of AI developments. The investment in public sector expertise must encompass an understanding of AI system development processes, deployment contexts, and ongoing operational characteristics. Unlike traditional regulatory subjects that remain relatively static once approved, AI systems continue evolving through training, deployment in new contexts, and interaction with other systems, requiring regulatory approaches that maintain ongoing oversight capability.[36]

Rather than simply regulating AI systems developed by private actors, public authorities can also actively shape AI development through strategic procurement, public research funding, and direct public sector AI development. These mechanisms enable public authorities to maintain technical expertise, influence development priorities, and demonstrate alternative approaches to AI system design and deployment.[37] State-led innovation in some sectors suggests that public authorities can play productive roles in shaping technological development when they maintain adequate capabilities and strategic vision.[38]

It is also important that institutional experimentalism enables productive state-market cooperation while maintaining clear boundaries around public authority and democratic accountability, without capture. Close collaboration between regulators and private participants can prove productive when properly structured, but requires careful attention to conflicts of interest, information asymmetries, and power dynamics.[39] AI supervisory authorities must establish governan-

[35] Kattel, R. et al., How to make an entrepreneurial state: why innovation needs bureaucracy (Yale University Press 2022).
[36] OECD, ‚Regulatory sandboxes in artificial intelligence' (2023) OECD Digital Economy Papers, No. 356.
[37] Mazzucato, M. & Beslon, C., L'État entrepreneur : pour en finir avec l'opposition public-privé (Fayard 2020).
[38] Anu Bradford, ‚The False Choice Between Digital Regulation and Innovation' (2024) 119(2) Northwestern University Law Review 402.
[39] Oskar Josef Gstrein and Anne Beaulieu, ‚Regulatory sandboxes in the AI Act: reconciling innovation and safety?' (2023) 49 Computer Law & Security Review 382.

ce mechanisms that enable technical collaboration while maintaining regulatory independence and public interest orientation. This requires transparent processes, conflict of interest management, and systematic evaluation of whether collaborative arrangements serve broader public interests rather than narrow industry preferences.[40]

Furthermore, AI systems' global nature creates complex challenges for domestic regulatory authorities seeking to maintain policy autonomy while enabling international coordination. Institutional experimentalism must include mechanisms for international regulatory coordination while preserving domestic democratic accountability and policy flexibility. The challenge becomes particularly acute when dealing with global technology firms that may have greater resources and technical capabilities than domestic regulatory agencies. Supervisory authorities must develop strategies that enable effective oversight of global AI systems while maintaining domestic policy autonomy and democratic accountability.

Implementation challenges and practical solutions

Translating institutional experimentalism from conceptual framework to operational reality requires systematic attention to implementation challenges that can undermine even well-designed experimental governance initiatives. Regulatory agencies possess entrenched organisational cultures, professional norms, and operational procedures that may resist experimental governance requirements. Successful implementation demands systematic change management that addresses both individual and organisational resistance.

First, change management must begin with clear sensitisation and communication about experimental governance rationale and its relationship to broader institutional missions. Civil servants need to understand how experimental activities serve institutional objectives rather than undermining traditional approaches. This requires leadership commitment that extends beyond policy statements to include resource allocation, performance evaluation, and career advancement criteria. Traditional regulatory training proves inadequate for staff engaging with experimental governance initiatives requiring different skills, knowledge, and approaches. On the contrary, comprehensive professional development programmes combining technical education, regulatory innovation methodologies, and collaborative governance skills are more suitable for this type of experimentalism. Professional development must address both technical competencies and institutional processes. Civil servants require an understanding of AI technologies, but they also need skills for managing experi-

[40] European Securities and Markets Authority and European Insurance and Occupational Pensions Authority, FinTech: Regulatory Sandboxes and Innovation Hubs (JC 2018 74, 2018).

mental programmes, coordinating across organisational boundaries, and engaging with diverse stakeholder communities. These skills often differ substantially from traditional regulatory competencies.[41]

Secondly, many regulatory agencies operate under legal frameworks designed for traditional approaches that may not provide clear authority for experimental governance initiatives. Implementation may require legislative reform to establish explicit authorities and procedural frameworks. Legal framework development must address the unique characteristics of experimental governance while maintaining essential oversight and accountability functions. This requires careful analysis of which procedural requirements serve essential regulatory objectives, and which may be modified or streamlined for experimental purposes.

Thirdly, experimental governance initiatives require dedicated resources for specialised staff, technical infrastructure, monitoring systems, and evaluation processes. These resource requirements must be balanced against other institutional priorities while maintaining sustainable funding for ongoing activities. Resource planning must therefore address both immediate implementation costs and longer-term institutional development needs. Experimental governance that remains permanently under-resourced may fail to generate lasting institutional capabilities or regulatory improvements, undermining its long-term effectiveness.

Fourthly, traditional regulatory performance metrics may prove inappropriate for experimental governance, which involves acceptable failures, iterative learning, and long-term outcomes that cannot be immediately measured. Supervisory authorities must develop strategic evaluation frameworks that capture learning and capacity-building benefits while maintaining accountability.[42] Performance measurement ought to balance accountability requirements with adequate weight for learning objectives. Experimental governance that becomes subject to traditional regulatory performance standards may lose its innovative capacity, while programmes that lack adequate accountability may lose institutional support and public legitimacy.[43]

Lastly, experimental governance must maintain appropriate standards while accommodating necessary flexibility and innovation. This requires quality assurance mechanisms specifically designed for experimental contexts rather than adapted from traditional regulatory oversight approaches. Quality assurance will ideally include peer review processes, external oversight mechanisms, and systematic evaluation procedures that identify successful practices and problematic outcomes.

[41] David A Wolfe, ‚Experimental Governance: Conceptual approaches and practical cases' (OECD Local Economic and Employment Development Papers, OECD Publishing 2020).
[42] Zachary Dove, Sikina Jinnah and Shuchi Talati, ‚Building capacity to govern emerging climate intervention technologies' (2024) 12(1) Elementa: Science of the Anthropocene 00124.
[43] David A Wolfe, ‚Experimental Governance: Conceptual approaches and practical cases' (OECD Local Economic and Employment Development Papers, OECD Publishing 2020).

Conclusion

Brazil's regulatory sandbox experiences offer a few lessons for AI supervisory authorities contemplating experimental governance approaches. Regulatory sandboxes, while potentially valuable for innovation governance, cannot function as universal remedies that resolve complex institutional and technological challenges without sustained investment in organisational change and capacity building. Nevertheless, the institutional experimentalism framework developed here addresses the organisational realities that often undermine experimental governance initiatives. Rather than treating bureaucratic resistance as an obstacle to circumvent, this approach recognises civil servants' concerns as legitimate expressions of institutional values that must be addressed through systematic design rather than dismissed or overcome.

The central insight proves particularly relevant for AI supervision: effective governance of emerging technologies requires active state engagement in shaping development trajectories rather than passive accommodation of private sector experimentation. This lesson gains urgency in AI contexts, where market concentration, societal implications, and democratic values create stakes that extend far beyond immediate innovation benefits.

AI supervisory authorities face a fundamental choice between two approaches to experimental governance. They can follow the path of minimal institutional change, treating sandboxes as temporary accommodations for private sector innovation while maintaining traditional regulatory approaches. Alternatively, they can embrace institutional experimentalism's more demanding approach of systematic capacity building, organisational change, and sustained investment in public sector capabilities. The first approach, while requiring fewer immediate resources and less institutional disruption, ultimately proves less effective at building sustainable capabilities for emerging technology governance. Regulatory sandboxes that remain isolated from broader institutional development risk becoming symbolic exercises that accommodate private-sector innovation without strengthening public-sector oversight capabilities. The second approach demands greater institutional commitment but offers the prospect of building lasting capabilities that extend beyond specific experimental initiatives. Institutional experimentalism requires sustained leadership commitment, systematic resource allocation, and willingness to undertake difficult organisational change processes. Yet the Brazilian evidence suggests that agencies pursuing this approach develop a more sophisticated understanding of emerging technologies, more effective stakeholder engagement capabilities, and more adaptive regulatory frameworks.

For AI governance, the stakes of this choice prove particularly high. The societal implications of AI systems, combined with rapid techno-

logical development and significant market concentration, create governance challenges that demand sophisticated institutional responses. Supervisory authorities that fail to build adequate capabilities risk becoming dependent on private sector assessments, unable to identify emerging risks, or incapable of protecting public interests in rapidly evolving technological landscapes. The path forward requires recognising that experimental governance represents not a departure from traditional regulatory values, but their adaptation to new technological and institutional contexts. The legal certainty, risk management, and public accountability that characterise effective regulation remain essential in experimental contexts, but they must be pursued through institutional approaches that accommodate uncertainty, enable learning, and maintain democratic oversight of technological development.

Successful governance of emerging technologies depends not on regulatory flexibility alone, but on institutional capabilities that enable ongoing adaptation while preserving essential regulatory functions. AI supervisory authorities that invest in building these capabilities position themselves to shape technological development in the public interest, while those that rely on minimal accommodation approaches risk becoming passive observers of privately determined technological trajectories. The challenge ahead is transforming experimental governance from an exceptional accommodation for private-sector innovation into a systematic capability for public-sector engagement with emerging technologies. Will it be easy? Absolutely not. Will it be worth it? The evidence from Brazil suggests yes, but only if institutions are willing to do the hard work of genuine transformation. This transformation requires sustained commitment to institutional development that extends far beyond individual sandbox initiatives, but offers the prospect of regulatory institutions capable of governing AI in the public interest while fostering beneficial innovation.

Disclaimer

The views expressed in this article are solely those of the author in his academic capacity and do not reflect the official position of the Agência Nacional de Proteção de Dados (ANPD - Brazilian Data Protection Authority). This independent research should not be construed as institutional endorsement of any regulatory approach discussed herein. Any errors remain the author's responsibility. This analysis is intended to contribute to academic discourse on AI governance and should be evaluated within that scholarly context.

References

Books

Bradford A, “The False Choice Between Digital Regulation and Innovation” (2024) 119(2) Northwestern University Law Review 402.

Kattel R, Drechsler W and Karo E, How to Make an Entrepreneurial State: Why Innovation Needs Bureaucracy (Yale University Press, 2022).

Mazzucato M and Collington D, The Big Con: How the Consulting Industry Weakens Our Businesses, Infantilizes Our Governments and Warps Our Economies (Allen Lane 2023).

Mazzucato M and Beslon C, L,État entrepreneur : pour en finir avec l‘opposition public-privé (Fayard 2020).

Journal articles

Allen HJ, “Regulatory sandboxes” (2019) 87 George Washington Law Review.

Chatzipanagiotou M and Koulierakis I, “AI Governance in a Complex and Rapidly Changing Regulatory Landscape: A Global Perspective” (2024) 11 Humanities and Social Sciences Communications 1134.

Cheng S-Y and Hou H, “Innovation, financial development, and growth: evidences from industrial and emerging countries” (2022) 55(3) Economic change and restructuring 1629.

Crampes C and Estache A, “Efficiency vs. distributional concerns in regulatory sandboxes” (2025) Journal of Economic Policy Reform 1.
Dove Z, Jinnah S and Talati S, “Building capacity to govern emerging climate intervention technologies” (2024) 12(1) Elementa: Science of the Anthropocene 00124.

Gstrein OJ and Beaulieu A, “Regulatory sandboxes in the AI Act: reconciling innovation and safety?” (2023) 49 Computer Law & Security Review 359.

Gstrein OJ and Beaulieu A, “Regulatory sandboxes in the AI Act: reconciling innovation and safety?” (2023) 49 Computer Law & Security Review 382.

Just SN and others, “Open for business: the discursive diffusion of regulatory sandboxes for fintech innovation” (2024) 17(3) Journal of Cultural Economy 360.

Lima CM de and Pasqualetto A, “Installation of Regulatory Sandbox Environment from the Perspective of the Brazilian Charter for Smart Cities” (2023) 11(83) Revista Nacional de Gerenciamento de Cidades.

Occasio W, Loewenstein J and Nigam A, “Institutional Logics: Motivating Action and Overcoming Resistance to Change” (2023) 19(6) Management and Organization Review 1171.

Quatrochi G and others, “Banks 4.0 in Brazil: possibilities to ensure fintechs financing role through its market positioning” (2023) 13(3) Innovation and Development 561.

Vieira JB and others, “O sandbox regulatório como instrumento de incentivo à inovação no Brasil: os casos do Banco Central do Brasil, da Comissão de Valores Mobiliários e da Superintendência de Seguros Privado” (2024) 153(1) Revista do TCU 343.

Washington PB and Lee E, “Nexus between regulatory sandbox and performance of digital banks: A study on UK digital banks” (2022) 15(12) Journal of Risk and Financial Management 1.

Yue L, Asgari B and Hawn O, “Regulation and Innovation Revisited: How Restrictive Environments Can Promote Destabilizing New Technologies” (2024) 35(4) Organization Science 1449.

Reports and working papers

European Securities and Markets Authority and European Insurance and Occupational Pensions Authority, FinTech: Regulatory Sandboxes and Innovation Hubs (JC 2018 74, 2018).

OECD, Regulatory Sandboxes for Privacy – Analytical Report (Business at OECD (BIAC) 2020).

OECD, “Regulatory sandboxes in artificial intelligence” (2023) OECD Digital Economy Papers, No. 356.

World Bank, Global experiences from regulatory sandboxes, Finance, Competitiveness & Innovation Global Practice (Fintech Note No. 8, International Bank for Reconstruction and Development 2020).

Wolfe DA, “Experimental Governance: Conceptual approaches and practical cases” (OECD Local Economic and Employment Development Papers, OECD Publishing 2020).

Government documents and official publications

ANPD, “ANPD’s Call for Contributions to the regulatory sandbox for artificial intelligence and data protection in Brazil is now open” (3 October 2023).

ANPD, “ANPD‘s Regulatory Sandbox Featured at report of the World Economic Forum” (26 January 2024).

ANPD, “ANPD prorroga prazo de inscrições para o sandbox regulatório de inteligência artificial e proteção de dados” (Autoridade Nacional de Proteção de Dados, 7 August 2025).

ANPD, “Edital nº 2/2025: Participação em Piloto de Ambiente Regulatório Experimental em Inteligência Artificial e Proteção de Dados” (Diário Oficial da União, 27 June 2025).

Banco Central do Brasil, Relatório de Gestão - Sandbox Regulatório: 1º Ciclo 2022 (Banco Central do Brasil 2022).

Comitê Estratégico de Gestão do Sandbox BC (CESB), Ata 1ª Reunião extraordinária do Comitê Estratégico de Gestão do Sandbox Regulatório (26 March 2021).

Comitê Estratégico de Gestão do Sandbox BC (CESB), Ata 2ª Reunião ordinária do Comitê Estratégico de Gestão do Sandbox Regulatório (16 July 2021).

Comitê Estratégico de Gestão do Sandbox BC (CESB), Ata 4ª Reunião ordinária do Comitê Estratégico de Gestão do Sandbox Regulatório (17 September 2021).

Online sources and other materials

Anjos L, “A bird’s-eye view of the Paris AI Action Summit: Regulation, power, and alternatives” Tech Global Institute (25 February 2025).
“Data Protection & Privacy 2024 – Brazil” (Chambers and Partners 2024).

REGULATORY SANDBOXES FOR AI AS TOOLS FOR TRUST: FROM CONCEPT TO PRACTICE

Author: Lorrayne Porciuncula; Datasphere Initiative

Abstract

Over the past two years, the use of regulatory sandboxes for artificial intelligence (AI) systems has gained significant traction.[1] This growing interest is driven by a set of well-founded goals: reducing information asymmetry between regulators and regulated entities; adopting agile regulatory tools capable of keeping pace with the rapid evolution of AI technologies; and developing responses that appropriately balance innovation with risk management, particularly in a domain where the societal and economic impacts are still unfolding.

Today, the central question is no longer whether sandboxes can be useful for addressing the regulatory challenges posed by AI, but rather how such novel tools should be designed and implemented to ensure they are fit for purpose and capable of delivering meaningful results.

What the most recent research on sandboxes reveals is that one of the critical factors for the success of such initiatives is the establishment of trust in the process. Although the definition of a sandbox may vary considerably depending on the sector in which it is applied, it is consistently described as a "safe space" for experimentation. However, trust in this process is not automatically granted by simply labelling it a sandbox. Building trust requires deliberate design choices and institutional practices that go beyond the nominal use of the term. In an era where trust matters as much as compliance, authorities must move beyond compliance and box-checking to design agile oversight systems that are transparent, inclusive, and adaptable. Thoughtfully planned and designed sandboxes can serve as critical spaces to cultivate that trust by enabling safe experimentation, meaningful engagement, and evidence-based learning.

This paper adopts an analytical and practice-oriented approach to examine the key design features of regulatory sandboxes that can support the relevant public entities and regulators in fostering trust, particularly in the context of AI governance.

These recommendations are grounded in an extensive, three-year research effort conducted by the Datasphere Initiative. As part of this work, we mapped and analysed over 60 data and AI-related sandbox initiatives at various stages of development across different jurisdictions worldwide.

[1] Datasphere Initiative, (2025), Sandbox for AI: Tools for a New Frontier, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai/; OECD (2023), "Regulatory sandboxes in artificial intelligence", OECD Digital Economy Papers, No. 356, OECD Publishing, Paris, https://doi.org/10.1787/8f80a0e6-en.

The analysis is structured around five core dimensions that were identified as critical to building trust in the sandbox process. First, it explores how to determine which emerging tech regulatory challenges are suitable for sandbox treatment and which require a traditional regulatory approach. The initiation phase is considered here as a crucial step required to define the sandbox's scope, objectives, and regulatory framing. Second, it addresses the involvement and engagement of cross-regulatory agencies, emphasising the need for a robust and coordinated governance structure. Third, it examines the design of the testing plan, including the formulation of hypotheses, the selection of appropriate methodologies, and the implementation of risk mitigation strategies. Fourth, it highlights the importance of developing a comprehensive communication and stakeholder engagement strategy to promote transparency, enhance legitimacy, and ensure the active participation of relevant actors throughout the process. Finally, it considers the establishment of clear evaluation criteria to assess the outcomes and overall impact of the sandbox initiative.

Introduction

Regulatory sandboxes are collaborative testing environments where regulators oversee the exploration of new technologies and business models.[2] By bringing together regulators, companies, and other stakeholders in time-bound processes, sandboxes reduce regulatory uncertainty and foster innovation through proactive and collaborative regulation.

Over the past two years, the use of regulatory sandboxes for artificial intelligence (AI) systems has gained significant traction, with the number of initiatives reaching over 60 examples by early 2025 (Figure 1).[3] A mapping by the Datasphere Initiative completed in January 2025 uncovered 66 sandboxes related to data, AI, or technology worldwide. While most are concentrated in the Global North and operate primarily at the national level, regional developments in Africa indicate a growing but more targeted adoption of the approach, particularly within the financial technology (fintech) sector.[4] Kenya's ICT sandbox, managed by the Communications Authority, exemplifies a model designed to test a broad range of ICT innovations, including AI.[5] Most of the mapped sandboxes are regulatory in nature, established and managed by ministries, departments, and agencies responsible for ICT and innovation, economic affairs, or data protection, with a smaller number led by sector-specific regulators.

[2] Datasphere Initiative (2025), Sandboxes for Data: Creating Spaces for Agile Solutions Across Borders, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-data/
[3] Datasphere Initiative, (2025). Sandbox for AI: Tools for a New Frontier, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai/; OECD (2023), "Regulatory sandboxes in artificial intelligence", OECD Digital Economy Papers, No. 356, OECD Publishing, Paris, https://doi.org/10.1787/8f80a0e6-en.
[4] Datasphere Initiative (2025), Africa Sandboxes Outlook, available at https://www.thedatasphere.org/datasphere-publish/africa-sandboxes-outlook/
[5] Communications Authority of Kenya, Regulatory Sandbox, Homepage, (Accessed 31 October 2025)

This growing interest is driven by a set of well-founded needs: reducing information asymmetry between regulators and regulated entities; adopting agile regulatory tools capable of keeping pace with the rapid evolution of AI technologies; and developing responses that appropriately balance innovation with risk management, particularly in a domain where the societal and economic impacts are still unfolding.

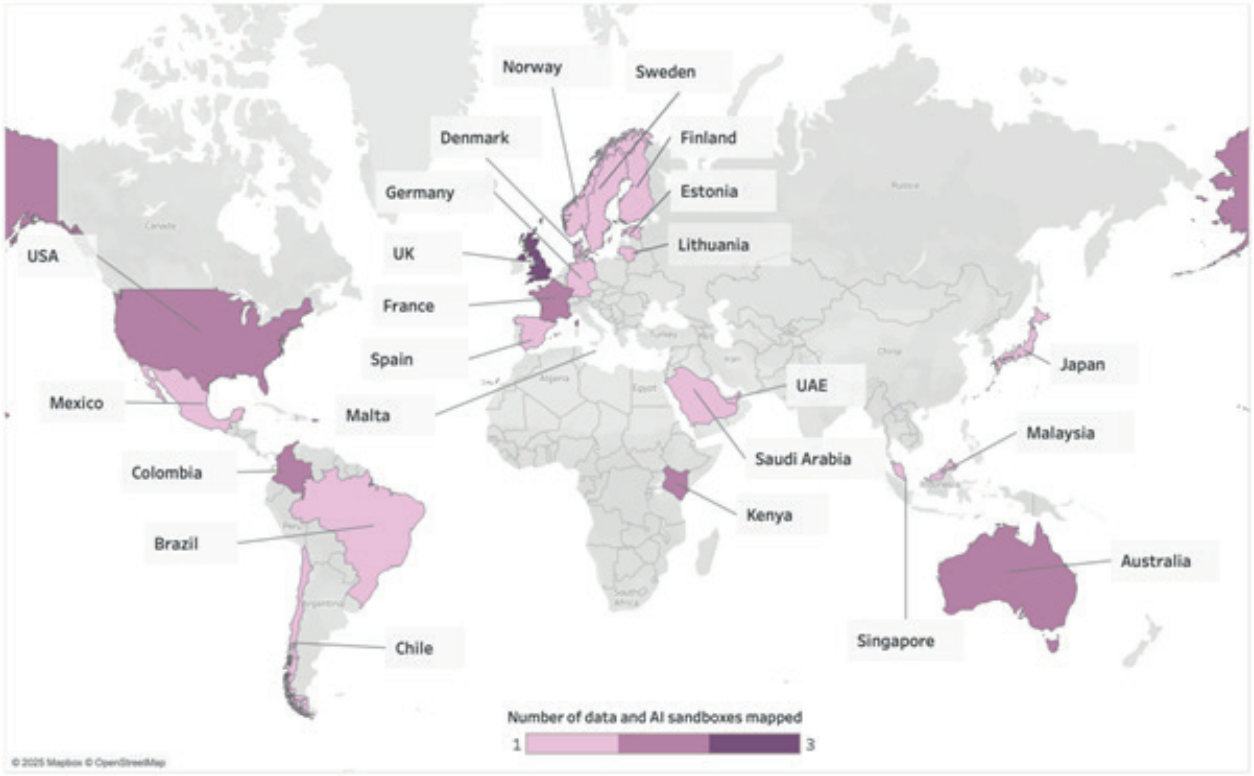

Figure 1. Map of Sandboxes for AI innovation around the world, as of January 2025
Source: Datasphere Initiative (2025), Sandboxes for AI: tools for a new frontier, available at: https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai/

In summary, AI regulatory sandboxes consist of controlled environments for iterative learning, stakeholder engagement, and real-world regulatory experimentation of AI products and services. With a growing number of sandboxes globally focusing on data and AI, these mechanisms are becoming essential tools for navigating emerging technology complexities. Examples have shown that AI sandboxes focus on different areas and goals, including:[6]

- Assessing emerging legal and technical issues relating to data protection and privacy: Sandbox cases here include Colombia's Superintendence of Industry and Commerce Sandbox, which focuses on privacy by design and by default in AI projects at the design stage involving personal data, and Estonia's Data Protection Panel/Sandbox by the Ministry of Economic Affairs and Communications, which targets projects that are priorities in the development of the digital state, particularly data processing that impacts fundamental rights.[7] [8]

- Testing compliance with data regulations: Example sandboxes here include the Norwegian Data Protection Authority Regulatory Sandbox for companies developing AI solutions in compliance with data protection regulations and the Swedish Authority for Data Protection's sandbox, which aims to identify "grey area" issues in their data protection law.[9] [10]

- Experimenting in order to identify and address trust issues: Examples here include Singapore's Generative AI Evaluation Sandbox for Trusted AI, led by the Infocomm Media Development Authority (IMDA), for evaluating and testing trustworthy GenAI, and Singapore's Privacy Enhancing Technology (PET) Sandbox, also by IMDA, which is essentially set up to increase trust in the use of privacy-enhancing technologies and to demonstrate their practical application in real-world projects.[11] [12]

Building on the experience of sandboxes in the financial sector, regulatory sandboxes for AI have been pushing the boundaries of facilitating product understanding for wider groups, gathering data to derive effective policies for evidence-based regulatory guidance, and ensuring compliance, public trust and accountability.

[6] Datasphere Initiative, (2025). Sandbox for AI: Tools for a New Frontier, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai.
[7] Colombia Superintendence of Industry and Commerce, Sandbox for privacy by design and by default in artificial intelligence projects, Homepage, (Accessed 31 October 2025)
[8] Estonia Ministry of Economic Affairs and Communications, The Digital Government Data Panel, Homepage, (Accessed 31 October 2025)
[9] Norwegian Data Protection Authority, Regulatory Privacy Sandbox, Homepage, (Accessed 31 October 2025)
[10] Swedish Authority for Privacy Protection, Regulatory Sandbox, Homepage, (Accessed 31 October 2025)
[11] The Infocomm Media Development Authority, Generative AI Evaluation Sandbox for Trusted AI, Homepage, (Accessed 31 October 2025)
[12] The Infocomm Media Development Authority, Generative AI Evaluation Sandbox for Trusted AI, Homepage, (Accessed 31 October 2025)

Challenges of AI Regulatory Sandboxes

Several factors contribute to the complexity of setting up AI sandboxes properly and earning the trust necessary for their success. The unique nature of AI creates several distinct challenges that must be addressed, including: the high stakes of AI, the gap between the technical and the regulatory world, the complex ecosystem of AI, and the global context of AI.

The high stakes of AI

AI regulatory sandboxes operate in an already sensitive and contested technological landscape. Unlike traditional sectors, where public scepticism develops gradually regarding emerging technologies, AI enters the regulatory experimentation space, carrying significant societal concerns about privacy, bias, job displacement, and algorithmic control.[13] These are not merely technical issues but fundamental questions about power, fairness, and societal values that sandbox processes must address transparently.

AI solutions are heavily data-driven and have complex, multidisciplinary elements, cutting across different areas and sectors. Their nature makes AI solutions prone to social conflict, ranging from data ownership and coordination issues to algorithmic bias and risks to people and the planet. A sandbox must not only evaluate a system‘s technical performance but also its broader societal impact and ethical implications

The gap between the technical and the regulatory world of AI

A significant challenge lies in the inherent complexity of AI. Many AI technologies often operate as “black boxes” that are difficult for both regulators and the public to understand or scrutinise. This creates a pressing need for a new level of AI knowledge among regulators to effectively assess AI products and services. Many regulatory authorities lack the technical expertise to understand AI systems deeply, creating a credibility gap that can undermine public confidence in the sandbox process. Users of these technologies and society in general need assurance that these experimental spaces are not simply providing regulatory cover for potentially harmful technologies. If not designed intentionally to foster trust, AI regulatory sandboxes can add a new layer of experimental and potentially opaque processes, causing the trust deficit to compound quickly.

[13] Sharma (2024), “Benefits or concerns of AI: A multistakeholder responsibility”, Futures, Vol. 157, available at https://doi.org/10.1016/j.futures.2024.103328

The complex ecosystem of AI

Sandboxes in more traditional sectors, usually highly regulated sectors, such as the financial sector, typically involve small stakeholder groups and are often regulated by a single authority. In contrast, AI sandboxes demand large, diverse stakeholder groups including multiple regulatory authorities, civil society organisations, rights groups, and the public. The practice of AI regulatory sandboxes necessarily needs to evolve into new ways to engage stakeholders and to understand emerging technologies and their associated risks and opportunities. This complexity means that a successful AI sandbox requires a robust and coordinated governance structure that can manage the diverse perspectives and goals of all parties involved, while ensuring the process is handled with the utmost transparency possible.

The global context of AI

While all regulatory sandboxes can benefit from some level of cross-border cooperation, AI sandboxes demand a higher degree of coordination. This is due to the inherently cross-jurisdictional nature of AI systems, which often rely on large, diverse datasets sourced across borders, and the evolving understanding of the technology itself. This global nature means that trust must be built not only locally but also across different regulatory cultures, legal frameworks, and public expectations. When a sandbox decision in one country affects AI development globally, the stakes for getting it right become even higher.

This evolution makes AI regulatory sandboxes complex policy experimentation tools to evaluate legal compliance, verify credibility and accountability, increase social acceptance, and mitigate regulatory risks while ensuring AI policy alignment globally. For success, transparency, trust, and inclusivity are critical elements that distinguish AI sandboxes from their other sectoral counterparts.

How to embed public trust in the AI regulatory sandbox journey

Sandboxes come in different formats and types, addressing different aspects of AI governance, which can create confusion and mistrust among stakeholders. There is a critical need to establish what elements can be standardised and which should be left to sector-specific, topic-specific, and jurisdictional contexts. When the public cannot easily discern how or why different sandboxes operate with varying structures and rules, a trust deficit can emerge and erode confidence in the process.

On the other hand, when set up properly with building trust as a foundational principle, AI sandboxes can help navigate these complexities, improve cross-border relations, and facilitate novel, innovative solutions that can greatly impact people and the planet. They have the potential to become dynamic spaces where public concerns are heard and addressed, where regulatory learning happens transparently, and where the benefits and risks of AI are evaluated collaboratively.

However, AI regulatory sandboxes are a new practice, and more work needs to be done in terms of identifying good practices specifically applicable to AI. The need for standardising approaches to sandboxing is not new, evidenced by efforts such as the OECD's regulatory sandbox toolkit, Datasphere Initiative's Global Sandbox Forum (and its upcoming Sandbox Assessment Framework), among other resources stemming from national experiences around the world.[14][15] The development of these tools demonstrates the importance of designing sandboxes effectively, not only to reap better outcomes but also to maintain public trust in the process.

Following structured frameworks like the Datasphere Initiative's 5-phase sandbox methodology (sandbox initiation, planning, execution, communication and engagement, and closure and evaluation) provides a vital roadmap (Figure 2).[16] However, each phase must explicitly and proactively integrate specific trust-building measures. Below, some key design features of regulatory sandboxes are shared in order to support relevant public entities and regulators in fostering trust, particularly in the context of AI governance.

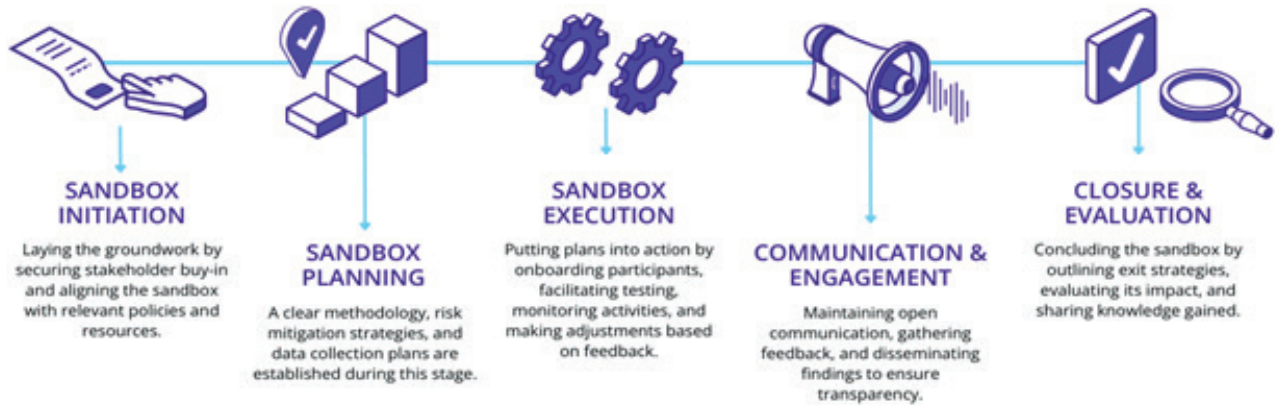

Figure 2. Structured framework for sandbox development phases
Source: Datasphere Initiative (2025), Sandboxes for AI: tools for a new frontier, available at: https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai/

[14] OECD (2025), Regulatory Sandbox Toolkit, available at https://www.oecd.org/en/publications/regulatory-sandbox-toolkit_de36fa62-en.html
[15] Datasphere Initiative (2025), "GSF Insights Session: Sandbox Assessment Framework #1", https://www.thedatasphere.org/news/global-momentum-builds-around-data-governance-experts-converge-for-high-level-sandbox-insights-session/
[16] Datasphere Initiative, (2025). Sandboxes for AI: Tools for a New Frontier, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai.

Sandbox initiation: trust-building grounded in a clear purpose and regulatory framing

The foundation of trust in AI regulatory sandboxes begins with clearly defining why the sandbox is necessary and whether the identified challenge is suitable for sandbox treatment versus traditional regulatory approaches. This critical first step establishes clarity and aligns expectations of stakeholders.

At this stage, teams must define the sandbox's purpose, scope, and objectives with particular attention to the crucial learning goals that justify its existence. Clear articulation of focus, duration, expected participants, and other design elements creates a strong foundation that builds stakeholder confidence in the initiative's legitimacy and potential effectiveness.

According to Brazil's 2025 Regulatory Sandboxes Reference Guide, the initiation phase helps regulatory authorities recognise areas requiring focused attention while establishing the goals, viability, and expected outcomes that will shape both immediate purpose and long-term strategic impact.[17] Teams must answer fundamental questions: What is the primary purpose of this regulatory sandbox? How do its objectives align with the authority's broader mandate? What specific regulatory, policy, or operational challenges justify experimentation?

Trust-building measures at this stage include:

- **Transparent problem definition:** clearly communicating what specific AI-related challenge the sandbox will address and why traditional regulation is insufficient
- **Stakeholder consultation:** engaging relevant parties early to validate the problem and proposed approach
- **Learning from peers:** drawing lessons from similar sandboxes to demonstrate informed decision-making and avoid known pitfalls
- **Clear success metrics:** defining measurable outcomes that allow public assessment of the sandbox's effectiveness

A thorough approach to initiation reduces design errors and builds early credibility by demonstrating that the sandbox is a thoughtful response to genuine regulatory challenges rather than an experimental exercise without clear justification. This transparency in purpose and process is essential for earning the trust necessary for successful AI governance experimentation.

[17] AGU/Labori and MDIC (2025), Regulatory Sandboxes Reference Guide, https://www.gov.br/agu/pt-br/regulatory-sandbox-reference-guide.pdf

Sandbox planning: intentional engagement to secure buy-in and build trust:

Trust-building during the planning stage centres on establishing transparent methodologies, comprehensive risk mitigation strategies, and inclusive stakeholder engagement processes. A clear methodology that defines internal processes, roles, and responsibilities throughout the sandbox lifecycle provides stakeholders with confidence in the sandbox's operational integrity. According to OECD's 2025 Regulatory Sandbox Toolkit, effective planning requires:[18]

- Identifying stakeholders and understanding their needs, interests, challenges, and uncertainties
- Categorising stakeholders into core participants, occasional contributors, and observers
- Establishing clear engagement criteria and communication flows

While this comprehensive approach requires more time to launch and additional staff resources to analyse feedback, it enables a better-defined approach tailored to specific contexts and builds stronger stakeholder confidence.

Moreover, a key trust-building element is meaningful involvement of civil society organisations (CSOs), which play a crucial role in ensuring AI governance remains inclusive, transparent, and accountable to public interests. A novel example here is the Institute for Future of Work (IFOW)'s Responsible AI Sandbox, which is a collaborative civil society-led sandbox engaging regulators, industry partners, and experts and offers active testing and evaluation of AI applications through a lens of fairness, accountability, and worker well-being.[19] CSOs help surface public concerns and integrate them into governance frameworks while ensuring underserved and underrepresented communities' perspectives are heard. Meaningful civil society participation in sandboxes requires intentional design choices, that is, building transparent processes – including public reporting on outcomes, clear documentation of lessons learned – as well as setting up structured engagement with affected communities – informing them about sandbox activities and inviting CSOs to participate in outreach events.[20] Ultimately, CSOs should be heard in decision-making processes and their input genuinely considered in sandbox design.

Thorough planning at this stage reduces the lead agency's sole responsibility and vulnerability to criticism while significantly increasing the sandbox's legitimacy and public trust. The investment in inclusive planning pays dividends in smoother execution and stronger stakeholder support throughout the sandbox lifecycle.

[18] OECD (2025), Regulatory Sandbox Toolkit, available at https://www.oecd.org/en/publications/regulatory-sandbox-toolkit_de36fa62-en.html
[19] IFOW, Institute for Future of Work (IFOW)'s Responsible AI Sandbox, Homepage, (Accessed 31 October 2025
[20] Datasphere Initiative, (2025). Sandboxes for AI: Tools for a New Frontier, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai.)

Sandbox implementation: ensuring trust through rigorous testing and risk mitigation

Trust during the implementation phase depends on demonstrating rigorous testing methodologies, robust data protection frameworks, and transparent risk management thoughtfully designed at the planning stage. This stage examines the design of the testing plan, including the formulation of clear hypotheses, selection of appropriate methodologies, and implementation of comprehensive risk mitigation strategies.

A critical trust-building element involves establishing clear protocols for confidentiality, data protection, and intellectual property rights. Teams must define how data will be collected, shared, stored, and disposed of securely, addressing public concerns about data misuse in AI systems. Transparent data handling procedures are essential for maintaining public confidence, particularly when sensitive personal or commercial data is involved in AI testing scenarios.

The methodology for the testing phase requires careful refinement of the core problem, identification of specific data types to be collected, alignment with learning objectives, and a clear specification of participant characteristics and requirements. Key design elements that inform the implementation phase of the sandbox include:

- Defining clear goals for each testing phase to ensure focused learning outcomes
- Determining appropriate testing environments, whether virtual settings, real-world deployments, or controlled simulations
- Outlining practical test operations, including structured interactions between participants, supervisors, and other stakeholders
- Outlining practical test operations, including structured interactions between participants, supervisors, and other stakeholders

Building on the planning stage, implementation requires operationalising risk mitigation strategies with a specific focus on AI-related concerns. Teams must identify and categorise potential risks, assess their likelihood and impact, and design appropriate response mechanisms with clear governance structures.

Critical risks to address include technical failures that could harm individuals or systems, data breaches or privacy violations, algorithmic bias that perpetuates discrimination, and regulatory misunderstandings. A significant risk involves perceived analogies where stakeholders misunderstand that participation in the sandbox might offer a form of regulatory leniency rather than rigorous oversight.

Effective implementation demonstrates that the sandbox operates with the same rigour as traditional regulation while providing the flexibility needed for innovative AI governance approaches. The goal is to build confidence that experimental regulatory approaches maintain high standards of protection and accountability while enabling valuable learning about AI governance challenges.

Communication and engagement: building trust through transparency

Trust in AI regulatory sandboxes requires continuous, strategic communication aimed at promoting transparency, enhancing legitimacy, and ensuring active participation of relevant actors throughout the process. A comprehensive communication and stakeholder engagement strategy becomes essential for maintaining public confidence in what can often appear as opaque regulatory experimentation.

Sustained trust requires continuous transparency about testing progress, interim findings, and any adjustments to methodology or risk mitigation approaches. This includes regular reporting to stakeholders, clear documentation of decisions and rationale, and accessible mechanisms for addressing emerging concerns during the testing phase. For AI sandboxes, this transparency is particularly crucial given public anxieties about algorithmic decision-making and data use.

A good practice around sandboxes involves designing, executing and monitoring communication strategies in each sandbox phase. This involves producing comprehensive reports and supporting materials that share key findings, best practices, and lessons learned, then disseminating these across the innovation ecosystem to inform future sandbox efforts and policymaking.

Effective communication serves multiple trust-building purposes: securing feedback and input, building stakeholder buy-in, and establishing regulatory relevance and credibility. Key communication elements include:

- Sharing the key problem the sandbox will address, along with the legal framework informing it, with core stakeholders to ensure understanding of the regulatory challenge and approach
- Communicating the objectives and goals of the sandbox with broader stakeholder groups once clarified to build collective understanding and support
- Clearly communicating about why innovators are participating and what they expect to achieve, helping public understanding of the sandbox's value
- Developing content with simple, accessible language to ensure the general public without technical expertise can understand sandbox activities, progress, and implications (e.g. social media campaign, blogs, podcasts, etc.)

Intentional data collection and documentation of sandbox activities, approaches, engagement outcomes, arising questions, and decision-making procedures enable effective communication throughout the journey. This systematic documentation serves as the evidence base for transparent reporting and accountability, allowing stakeholders to track how decisions were made and problems were addressed.

For AI governance, this documentation becomes particularly important as it creates an audit trail that can address concerns about regulatory capture or insufficient oversight of powerful technologies. The communication strategy must balance transparency with protection of legitimate commercial interests and sensitive regulatory information, demonstrating that openness and responsible governance can coexist.

Evaluation and closure: demonstrating impact and sustaining trust

The final stage of building trust through AI regulatory sandboxes involves establishing clear evaluation criteria to assess both the outcomes and overall impact of the sandbox initiative. Rigorous evaluation demonstrates accountability to all stakeholders and provides evidence that the experimental approach generated meaningful insights worth the investment of public resources and regulatory attention.

Comprehensive evaluation serves multiple trust-building functions by providing different stakeholders with the information they need to assess the sandbox's value and impact:

- **Participating companies** receive guidance through detailed "Exit Reports" that document their testing experience, regulatory insights gained, and pathways forward for their solutions
- **Regulatory authorities** obtain guidelines and recommendations through "Impact Assessment Reports" that function as regulatory prototypes, informing future policy development and governance approaches
- **The broader public and affected communities** access "Final Reports" on the sandbox program that demonstrate transparency and accountability in how their interests were protected and advanced

The evaluation draws on regular monitoring and reporting carried out during the testing phase, ensuring that the assessment is grounded in systematic data collection rather than retrospective impressions. Key evaluation activities include:

- Ensuring transparency and clarity in both processes and results through systematic documentation

- Identifying and assessing companies' engagement levels and outcomes throughout the testing process
- Conducting structured interviews with all participants, including companies, regulators, and stakeholder representatives
- Engaging with the broader public and stakeholder communities established in previous stages to gather their perspectives on sandbox impact

Evaluation covers both the sandbox setup and operations as well as the substantive outcomes, examining results alongside process considerations. This dual focus ensures that lessons are learned not only about the AI technologies tested but also about the sandbox methodology itself, contributing to continuous improvement in regulatory experimentation approaches.

The evaluation stage ultimately determines whether the sandbox succeeded in generating valuable regulatory learning, protecting public interests while enabling innovation, operating with sufficient transparency and accountability and ultimately shaping both immediate policy decisions and the future credibility of regulatory sandboxes as tools for AI governance.

Conclusion

The evolution of AI regulation presents a fundamental challenge: how to govern rapidly advancing technology, without falling into a dichotomic choice between innovation vs responsibility, in a way that is both agile and trustworthy. As argued above, regulatory sandboxes offer a powerful answer, but only if their design is deliberately and systematically focused on building and maintaining public trust. These sandboxes are not simply a modern alternative to traditional regulation; they are a vital space for collaborative learning and a crucial vehicle for restoring public confidence in governance during a period of profound technological change.

The framework presented here, grounded in a five-phase approach, is a step in that direction. The journey begins with Sandbox Initiation, where clear purpose and transparent framing establish the initiative's legitimacy and set the stage for success. This is followed by a meticulous Planning phase that secures broad buy-in by meticulously identifying and integrating a diverse array of stakeholders, especially civil society organisations. As the sandbox moves into Implementation, trust is solidified through rigorous testing, robust risk mitigation, and a commitment to data protection. This is continuously reinforced by a proactive and open Communication and Engagement strategy that informs all phases, in order to ensure stakeholders and the

public are informed at every step. Finally, the sandbox concludes with a rigorous Evaluation and Closure phase, providing the evidence and accountability needed to validate the process and demonstrate its tangible impact on policy and society.

By systematically applying these steps, regulators can address the core challenges that threaten the credibility of AI governance and AI regulatory sandboxes. The lack of technical expertise among regulators can be mitigated through the direct, hands-on learning provided by the sandbox, which allows them to understand the nature of these systems and the incentives of the stakeholders behind them. The high stakes and social conflicts inherent to AI can be managed through the meaningful inclusion of diverse stakeholders, ensuring that ethical concerns and public anxieties are addressed proactively, not as an afterthought. Furthermore, the complex, multi-jurisdictional nature of AI can be addressed by the framework when transparency and cross-border learning are taken seriously.

The true value of an AI regulatory sandbox is not measured by the number of companies it supports or the speed at which it operates, but by its ability to foster genuine, lasting trust. This trust is the essential currency that allows society to embrace the benefits of AI while effectively managing its risks. It serves as the bridge between technological innovation and public acceptance, ensuring that AI systems are developed and deployed in a manner that aligns with societal values and expectations. As a new era of AI unfolds, the success of a regulatory sandbox will be defined by its ability to act as a tool not just for compliance but for creating a shared foundation of confidence. It is through these collaborative, transparent, and accountable experiments that we can build a future where AI serves humanity in a way that is both innovative and secure.

This forward-looking perspective is crucial because AI governance will not be a one-time act but a continuous process of adaptation. Sandboxes are not a final solution but a dynamic, evolving methodology for navigating the future of technology. They provide a model for how public and private sectors can co-create policy and standards in real-time, moving beyond reactive regulation to a proactive and preventative stance. This approach helps build the institutional capacity to respond to future AI advancements, whatever they may be. By committing to the trust-building strategies outlined here – from clear purpose and inclusive planning to transparent communication and rigorous evaluation – authorities can ensure that regulatory sandboxes become more than just a passing trend. They can become a permanent, foundational element of an agile, ethical, and effective governance system for AI, paving the way for a trustworthy and responsible digital future.

References

1. Datasphere Initiative (2025), Sandbox for AI: Tools for a New Frontier, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai/; OECD (2023), "Regulatory sandboxes in artificial intelligence", OECD Digital Economy Papers, No. 356, OECD Publishing, Paris, https://doi.org/10.1787/8f80a0e6-en.

2. Datasphere Initiative (2025), Sandboxes for Data: Creating Spaces for Agile Solutions Across Borders, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-data/

3. Datasphere Initiative (2025). Sandbox for AI: Tools for a New Frontier, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai/; OECD (2023), "Regulatory sandboxes in artificial intelligence", OECD Digital Economy Papers, No. 356, OECD Publishing, Paris, https://doi.org/10.1787/8f80a0e6-en.

4. Datasphere Initiative (2025), Africa Sandboxes Outlook, available at https://www.thedatasphere.org/datasphere-publish/africa-sandboxes-outlook/

5. Communications Authority of Kenya, Regulatory Sandbox, Homepage, (Accessed 31 October 2025)

6. Datasphere Initiative (2025). Sandbox for AI: Tools for a New Frontier, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai.

7. Colombia Superintendence of Industry and Commerce, Sandbox for privacy by design and by default in artificial intelligence projects, Homepage, (Accessed 31 October 2025)

8. Estonia Ministry of Economic Affairs and Communications, The Digital Government Data Panel, Homepage, (Accessed 31 October 2025)

9. Norwegian Data Protection Authority, Regulatory Privacy Sandbox, Homepage, (Accessed 31 October 2025)

10. Swedish Authority for Privacy Protection, Regulatory Sandbox, Homepage, (Accessed 31 October 2025)

11. The Infocomm Media Development Authority, Generative AI Evaluation Sandbox for Trusted AI, Homepage, (Accessed 31 October 2025)

12. The Infocomm Media Development Authority, Privacy Enhancing Technology Sandboxes, Homepage, (Accessed 31 October 2025)

13. Sharma (2024), "Benefits or concerns of AI: A multistakeholder responsibility", Futures, Vol. 157, available at https://doi.org/10.1016/j.futures.2024.103328

14. OECD (2025), Regulatory Sandbox Toolkit, available at https://www.oecd.org/en/publications/regulatory-sandbox-toolkit_de36fa62-en.html

15. Datasphere Initiative (2025), "GSF Insights Session: Sandbox Assessment Framework #1", https://www.thedatasphere.org/news/global-momentum-builds-around-data-governance-experts-converge-for-high-level-sandbox-insights-session/

16. Datasphere Initiative (2025). Sandboxes for AI: Tools for a New Frontier, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai.

17. AGU/Labori and MDIC (2025), Regulatory Sandboxes Reference Guide, https://www.gov.br/agu/pt-br/regulatory-sandbox-reference-guide.pdf

18. OECD (2025), Regulatory Sandbox Toolkit, available at https://www.oecd.org/en/publications/regulatory-sandbox-toolkit_de36fa62-en.html

19. IFOW, Institute for Future of Work (IFOW)'s Responsible AI Sandbox, Homepage, (Accessed 31 October 2025)

20. Datasphere Initiative (2025). Sandboxes for AI: Tools for a New Frontier, https://www.thedatasphere.org/datasphere-publish/sandboxes-for-ai.

SECTION III.

INNOVATION THROUGH SUPERVISION

If the previous sections explored how authorities can build interpretative and experimental capacities, this part of the report turns to why such capacities matter: because supervision, when done well, can become one of the most powerful engines of innovation. This idea was at the centre of the third session of the Expert Roundtable, where panellists argued that well-designed oversight builds the very trust and predictability that innovative markets need to flourish.

The conversation emphasised that effective supervision is not about restraint but about creating conditions for responsible risk-taking. Supervisors who engage transparently with stakeholders, acknowledge uncertainty, and learn iteratively from real-world cases help shape markets that reward integrity and social value rather than regulatory arbitrage. In this sense, supervision can guide innovation toward outcomes that are fair, inclusive, and societally beneficial.

Authorities are already demonstrating that effective supervision does not slow innovation down but gives it direction. Oversight creates the conditions in which companies compete on quality, ethics, and public value rather than on exploiting regulatory blind spots. When rules are clear, consistently applied, and embedded in transparent engagement between supervisors and the market, they enable risk-taking that is responsible rather than reckless.

Building this kind of environment requires investment in culture as much as in capability. As markets evolve rapidly, supervisors must be empowered to act flexibly, drawing on multidisciplinary teams and shared expertise. Open communication, even around challenges or setbacks, strengthens both legitimacy and learning. Rather than trying to "catch up" with technology, supervisors can lead by example, showing that regulation grounded in dialogue and evidence can be as adaptive as the systems it governs.

The following contributions illustrate how this vision is taking shape in practice. Håkan Burden and Susanne Stenberg examine how collaborative and team-based supervision models can cultivate trust and innovation readiness within public institutions. Their analysis is complemented by another perspective exemplifying how transparent, iterative, and well-resourced oversight can help supervisors guide innovation responsibly while earning the confidence of those they supervise.

BUILDING AI CAPACITY IS TEAMWORK: EXPERIENCES FROM EXPLORING THE UNKNOWN TOGETHER

Authors: Håkan Burden, Susanne Stenberg; Rise Sweden

Abstract

This contribution examines how public authorities can build supervisory capacity for AI by engaging directly in collaborative innovation processes. Drawing on two Swedish policy lab initiatives, Policylabb Smarta Fartyg (PLSF) and Policy Lab Urban Zjöfart (PLUZ), the paper demonstrates how co-creation, openness, and trust enable authorities, industry actors, and researchers to jointly interpret policy in the context of emerging AI-enabled maritime technologies. Rather than focusing on market surveillance or certification, the policy labs served as structured environments for exploring how existing policies apply to smart ships and Maritime Autonomous Surface Ships (MASS), and for refining regulatory guidance through both conceptual analysis and real-world trials. The findings highlight that building AI supervisory capacity is inseparable from understanding evolving business models, engaging in iterative learning, and developing shared safety practices such as the use of safety cases. The contribution argues that authorities can foster market maturity by participating in innovation ecosystems, maintaining transparency, and developing first-practice interpretations of policy that can evolve into widely adopted best practices. In doing so, it shows that effective AI supervision is not a solitary regulatory task but a collective endeavour across public and private stakeholders.

Introduction

Our aim with this contribution is to provide insights into how public authorities can build their supervisory capacity through initiatives facilitating innovation and thereby also market maturity. This is a topic that has seen growing attention lately, not least through the European Union's emphasis on regulatory sandboxes, with claims to foster innovation and regulatory compliance.[1] The contribution is primarily concerned with the role of the national authority and its capacity to supervise others' innovation.

In short, our message is that if authorities want to be involved in others' innovation of AI, it requires a different mindset than that of market surveillance:

- Openness - insights should be shared openly so all have an opportunity to learn,
- Co-creation - learning from each other and seeing new perspectives, and
- Trust - so uncertainties and possible mistakes can be addressed instead of obscured.

All are key factors as a first practice of applying regulation to innovation is sought.

Our contribution is grounded in the evaluation of two research projects we conducted on the regulatory compliance of novel AI systems within the maritime sector - Policylabb Smarta Fartyg (Policy Lab Smart Vessels, PLSF) and Policy Lab Urban Zjöfart (Policy Lab Urban Zhipping, PLUZ).[2] PLUZ). Both projects investigated existing policy in relation to the adoption of smart ships, or Maritime Autonomous Surface Ships (MASS[3]), which deploy AI for automating operations such as docking and cruise control.

We use the term policy instead of regulation or law since our definition encompasses both. By policy, we mean the intention to influence the actions of others for a specific purpose.[4] This intention can take the form of law, international standards, codes of conduct, guidelines and decisions from competent authorities, etc. When we conduct research on policy in relation to innovation, we refer to the methodology as a policy lab.

[1] Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, A Competitiveness Compass for the EU, Brussels, 29.1.2025 COM (2025) 30 final. Retrieved from: https://commission.europa.eu/document/download/10017eb1-4722-4333-add2-e0ed18105a34_en?filename=Communication_1.pdf
[2] Shipping spelt with a 'z' to distinguish it from other projects. It was a decision made just before the submission deadline.
[3] International Maritime Organisation, Autonomous shipping, https://www.imo.org/en/mediacentre/hottopics/pages/autonomous-shipping.aspx
[4] This resonates with Black's definition of regulation, see Black, J. (2022). Critical Reflections on Regulation, Australian Journal of Legal Philosophy.

For us, policy labs aim to clarify the application of policy in relation to innovative products and services, so that relevant stakeholders know who has what mandate to act. The methodology is inspired by design research in that we seek a shared understanding of the challenges at hand and the co-creation of solutions. This is, in turn, done iteratively, so that the problem understanding is enriched by the exploration of possible solutions, and vice versa. Our role in the process is both as a neutral facilitator and as impartial researchers, conducting our own data collection and analysis. That requires an understanding of research methodologies, innovation processes, policy and technology as well as access to domain experts. As researchers, we make no decisions relating to market surveillance nor business investments, that is for the pating parties to decide upon.

This contribution is structured as follows: The first two sections describe a policy lab each. The sections share the same structure,[5] detailing the organisation, purpose, operation and outcome of each policy lab. That enables comparisons across policy labs. In the first policy lab, Policylabb Smarta Fartyg, the focus was on understanding which policies enable the introduction of smart ships on Swedish waters and the implications on mandates to act. In the second policy lab, Policy Lab Urban Zjöfart, the focus was on trialling the smart ships and their AI systems together with the policies under real-world conditions. The lessons learned from reflecting on the experiences are summarised as insights and recommendations before we conclude with how authorities help foster market maturity by involvement in others' innovation.

Co-creating policy for novel AI systems: Policylabb Smarta Fartyg (PLSF)

Policylabb Smarta Fartyg (PLSF) was conducted from March 2021 to June 2022. The focus of the project was on exploring which policies are applicable in relation to the systems brought to the project by the participants. A major aspect in the process was establishing trust between the public authority and the private enterprises in sharing activities already taking place.

Organisation

The consortium behind the policy lab consisted of six partners – the Swedish Transport Agency, responsible for supervision of ships and maritime operations; the Swedish Maritime Administration, responsible for managing maritime infrastructure; Saab Kockum, a commercial technology developer; ABB, a commercial technology developer; Färjerederiet, a public body operating passenger traffic; and RISE Research Institutes of Sweden, a research institute.

[5] Burden, H., & Stenberg, S. (2025). An intermediate governance model for EU regulatory sandboxes. Retrieved from: https://diva-portal.org/smash/record.jsf?pid=diva2:1995583

The project was funded through the Swedish Transport Administration's research portfolio (the Swedish Transport Administration and the Swedish Transport Agency are two different authorities, where the former has a primary responsibility of providing land-based infrastructure, while the latter is responsible for supervision of mobility and transportation on land and sea, as well as in the air).

Purpose

The scope of PLSF was to explore regulatory aspects of smart ships and to provide initial guidance on introducing AI-based systems and services in the maritime sector. In neither of the projects was the implementation of the AI systems under investigation; it was their adoption in relation to national policies on maritime safety that was explored.

An important aspect throughout the project was that the involved authorities were invested because they wanted to understand the regulatory landscape in relation to the needs of providers and deployers of smart shipping solutions. Thus, their role was not market surveillance. By participating, the ambition was to clarify the regulatory landscape in relation to smart ships and their novel AI systems. This, in turn, relies on concrete details about the application and use of the systems, which were provided by the technology developers and Färjerederiet. RISE contributed with policy analysis, maritime expertise and project management.

Operation

The policy exploration emanated from three cases:

1. Remote pilotage carried out by the Swedisch Maritime Administration
2. Manned road ferries with AI-assisted docking and cruise control, operated by Färjerederiet, and
3. Autonomous and unmanned boats operated from shore or another boat, developed by Saab Kockums.

The first case did not involve AI systems, so it will not be further discussed.

PLSF was conducted as monthly meetings where we took turns to present analyses and discuss insights and open questions. This meant that both the authorities and the other partners contributed towards mapping out the policy landscape and assessing which policies were relevant for the three cases. This required details on the nature of the operations and the chosen technology. In the last months of the project, we formalised the insights by relating them to two conceptual

cases. This enabled a more detailed analysis while not disclosing trade secrets. During PLSF, the involved authorities were never put in a position where they needed to exercise their supervisory powers.

For PLSF, the main resources consisted of the funding from the Swedish Transportation Administration and the competencies of the involved partners. As the project was conducted during the pandemic, we relied heavily on a platform facilitating virtual workshops and file sharing. In this case, co-location was not a prerequisite for carrying out the project activities. The policy lab did not need any technical infrastructure for training, testing or validating AI systems since the aim was to distil the relevant policies and analyse the cases in relation to them.

Outcomes

The outcomes are described in a public project report and a conference contribution. The report details the aim, organisation, ways of working and the regulatory guidance on adopting smart ships and related AI systems in a Swedish context. [6] [7]

In short, the three policies were:

1. A written statement of intent between the Swedish Transport Agency and the Swedish Maritime Administration to start investigating the possibility of trialling remote pilotage, with the intent to ensure the shipping industry that they were collaborating on developing new services meeting an increased demand on waterways.
2. Certification of ferries is commonly done in relation to an established and consistent set of technical requirements. For smart ships such as the new road ferries, it is reasonable to complement traditional certification with a safety case to ensure that the ship is seaworthy.
3. Ships less than five meters long that do not carry passengers are excluded from national rules regulating the market surveillance performed by the Swedish Transport Agency. If the police or coast guard were to conduct an inspection, it would be necessary to show how the operations comply with responsibilities such as pollution and collision avoidance.

All three policies referred to the authorities' own directives, national law and international instruments. In this way, the outcome not only gives recommendations on how to move forward, but it also shows how policies at multiple levels of mandate interact to support the outcome. During the project, the Swedish Transport Agency came out with its own guidelines for trialling smart ships on Swedish waters.[8]

[6] Burden, H., Stenberg, S., Carlgren, L., & Sjöblom, T. (2022). Policylabb Smarta Fartyg. In Swedish but includes an English summary. Retrieved from: https://diva-portal.org/smash/get/diva2:1707210/FULLTEXT01.pdf
[7] Burden, H., Stenberg, S., Carlgren, L., & Sjöblom, T. (2023). The Swedish policy lab for maritime autonomous surface ships. Transportation Research Procedia, 72, 1840–1847. Retrieved from: https://doi.org/10.1016/j.trpro.2023.11.661
[8] The Swedish Transport Agency, Transportstyrelsens riktlinjer för tester med smarta fartyg, TSS 2020-4309 (Swedish only). Retrieved from: https://www.transportstyrelsen.se/globalassets/global/sjofart/autonom-sjofart/transportstyrelsens-riktlinjer-for-tester-med-smarta-fartyg.pdf

Real-world testing of AI systems and policy: Policy Lab Urban Zjöfart (PLUZ)

Policy Lab Urban Zjöfart (PLUZ) was carried out from February 2023 to February 2025. While the ambition was still to establish a way of introducing smart ships on Swedish waters in a reasonably safe manner, there was now a new consortium responsible for the exploration.

Organisation

In PLUZ, the partners were Torghatten, a private operator of passenger ferries; Zeabuz, a startup technology developer; Ports of Stockholm, responsible for the quays and waterways of Stockholm; DNV, a classification society providing rules and assurance for shipping; and RISE. The main reason for the Swedish Transport Agency not being a partner was a shortage of available staff. Again, the project was funded through the Swedish Transport Administration's research portfolio.

Purpose

The scope of PLUZ was subsequently to apply the guidance from PLSF on a concrete case. In this way, the two projects form a trajectory where the first project focused on delivering regulatory guidance for accessing the maritime market with AI-based solutions, while the second project focused on testing both the AI-based solutions and the policies under real-world conditions. It was never the intention of PLUZ to put the development of the AI systems under investigation; it was their adoption in relation to national policies on maritime safety and innovation that was explored.

There were two trials notified and conducted during the project:

1. What is the information loss when navigational instructions are transferred from the steering console to the engine room using wireless communication, compared to traditional wired solutions?
2. What lessons can be drawn when the ferry in Stockholm, Sweden, is remotely operated from Trondheim in Norway?

The second trial was not decided on before the first trial was evaluated. The results showed that while there was some loss in information, this was not critical due to the redundancy of information in the navigational signals.

Operation

For PLUZ, the submission of notifications for trialling smart ships, the trialling itself and an initial evaluation were the main phases and activities. As the relevant authorities were not part of the consortium, we chose to organise specific meetings and demonstrations to provide them the opportunity to see the systems at work and, together with the consortium, discuss learnings and open questions. We also kept the Swedish Transport Agency informed of our plans so that they had time to muster their internal resources for receiving a first-ever notification of trials of smart ships on Swedish waters. Receiving a notification also made it explicit how the authority was to process the notification and reply.

The certification of MF Estelle followed the second policy of PLSF, so that the decision was based on a combination of established principles and a safety case for the novel designs. This was done before the project launch and was not a project activity in itself. There is also an important distinction between the certification of MF Estelle and her crew, and the process for conducting trials, as the former is an example of when the authority must give an approval for a product to be on the market, while the latter is supervision of innovation and providing regulatory guidance.

In terms of resources, Torghatten contributed with their ferry MF Estelle, which included Zeabuz's autonomous systems. The ferry was operated on Riddarfjärden in Stockholm, while Zeabuz's remote operation centre, used for the second trial, was placed in Trondheim, Norway. Zeabuz kept a record of the conducted trials and automatically logged data from the trials on its own servers. We also reused the platform for virtual workshops and file sharing.

Outcomes

The importance of safety cases was raised during PLSF as an outcome of the regulatory analysis. In PLUZ, we developed two safety cases as attachments to the two notifications submitted to the authority. [9] The safety cases included cybersecurity aspects as well as passenger safety and were made public together with the notifications as part of the research project.[10]

Recommendations

Based on the account of the two policy labs (PLSF and PLUZ), we want to highlight four overarching insights:

[9] Due to the Swedish Governance principles, the notifications are public, and anyone can request access to the correspondence and the attached documents through the authority.
[10] Burden, H., Stenberg, S., Nilsson, E., & Petersson, C. (2025). Slutrapport Policy Lab Urban Zjöfart, with English summary. Retrieved from: Retrieved from: https://diva-portal.org/smash/get/diva2:1943322/FULLTEXT01.pdf

1. Regulatory guidance is a team effort,
2. Be transparent and open
3. Business development is as important as technical innovation, and
4. Keep it simple.

This requires trust among the participants. And trust takes time to develop, so ensure there is time and opportunities for getting comfortable with sharing uncertainties and perceived weaknesses among the consortium members.

Competent teams establish first practice

Market maturity requires competent authorities.[11] That requires an understanding of how the ecosystem (i.e., the market and its actors) is changing due to innovative solutions and new policies. Being involved in initiatives that facilitate innovation, such as regulatory sandboxes, is a hands-on way of understanding the possible implications of the changing landscape. Coming back to co-creation, the authority is not the master while the rest of the consortium are attentive novices waiting to be taught. It is rather a question of understanding the changes together.[12] The different roles and mandates of the actors will then have an impact on recommendations and developments - the authorities will have more impact on the application of policy, while private enterprises will make their decisions on which systems to develop and how.[13]

The authority needs to understand the system and its context to some degree, just as the perspective of the innovating parties on policy highlights new interpretations and a lack of clarity. Together, the parties will establish a first practice where there is none – "this is how we interpret the policy in the light of new circumstances, and this is how we chose to act on that knowledge".

Success can then be seen as the establishment of (more) clarity in how to comply with new policies and how innovative systems change the interpretation of existing policy.[14]

Why transparent AI if authorities are opaque?

There was no need for a strategy on the sharing of IP in either of the projects. Something that can be perceived as important and complicated was straightforward, as each partner was free to adopt what was relevant for their organisation. The know-how and information

[11] This is an intentional pun. Market maturity relies on authorities with the mandate to act, and the authorities need competent personnel to decide when and how to act.
[12] Burden, H., & Stenberg, S. (2023). Sustainable AI and Disruptive Policy – AI Regulatory Sandboxes. Retrieved from: https://ri.diva-portal.org/smash/get/diva2:1835556/FULLTEXT01.pdf
[13] We found the same team effort when we participated as observers in The Swedish Authority for Privacy Protection's project Disclosure of Public Records Using AI, see: The Swedish Authority for Privacy Protection (2024). Disclosure of Public Records Using AI, see https://www.imy.se/globalassets/dokument/rapporter/english-summary-disclosure-of-public-records-using-ai.pdf (English summary)
[14] This resonates with the EU regulations introducing regulatory sandboxes, see: Burden, H., & Stenberg, S. (2025). EU regulatory sandboxes - An opportunity for coordinating AI Innovation. Retrieved from: https://diva-portal.org/smash/record.jsf?pid=diva2:1995581

regarding technology that was shared among the participants was not of the nature of trade secrets, hence we did not need non-disclosure agreements (NDAs) or explicit decisions on the sharing of the project outcomes.

Avoiding opaque governance is also an important aspect of the purpose of innovation support. The products are not yet on the market, so supporting innovation means getting involved in research and product development. Hence, the role of the authority is not market surveillance but supervision in the sense of giving regulatory guidance, ensuring that recommendations are feasible from the perspective of the authority and respecting the agreements with project participants and overall policies governing the authority.

We therefore suggest adjusting the sharing of information to the purpose of the activity - if the objective is to facilitate innovation, it is not necessary to run a test suite on source code, document under which circumstances a specific cybersecurity solution can be hacked or certify the competence of employees.[15] Instead, the focus should be on relating test methodologies for an AI system to relevant policy requirements, how different cybersecurity solutions interact and balance the overall safety case, and how training and mandates match the responsibilities of the involved personnel. By sharing the insights from the latter, the whole ecosystem can benefit from the authorities' guidelines.[16]

An important ethical question in both projects was how to balance the impartiality of the relevant authorities while enabling them to get a deeper understanding of the technology and business models for smart ships. The risk of partiality must be managed by the involved authorities, but all involved parties must recognise the need and respect the balance struck by the authorities. Paying attention to the question of partiality throughout the project also facilitates clarity on roles and responsibilities within the project consortium. Being open and transparent is one way of maintaining impartiality; if an authority can openly share what they are doing, then there is a lower risk of giving specific actors favourable treatment.

The bottom line is business development

It is easy to get over-focused on the AI system at hand. We have found that while the uncertainty around policies and digital innovation is important to address, it is just as important to remember that the innovators will swiftly change track if they perceive that the innovation will have more impact in another setting or if another technical solution better fits the ambitions of the organisation.

[15] However, if both the authority and the other participants deem it important to explore these kinds of questions, the information sharing can be managed by conducting the testing using the toolchain already used by the system developer. What to share and to what detail is then under the control of the developer. This has the added benefit of not having to fit the system under development to the toolchain of the authority.
[16] Ensuring that the authority's guidance on first practice is suitable for a specific innovation requires innovators to bridge the gap between the abstract guidance and the concrete details of their systems. This holds for both the innovators participating in the initiative and those on the outside who adapt to the outcomes. In short, one needs to apply judgment even when collaborating with a competent authority.

As PLUZ was initiated, the idea was to explore to what extent the master of the ship could operate the vessel from a remote operation centre onshore. As new experiences of the system were gained through the trials, there was a shift in perceived value so that the value of the AI system is to be realised by reducing the number of personnel and/or the needed competence onboard other vessels, as some roles or responsibilities can be managed by the AI systems or from shore.
So, while the project might have the innovative AI system as its focal point, the deployers of the system will have the business opportunities in their long-term sight. In parallel, for the authority, it is the learnings in relation to how policies can be applied to tomorrow's products and the implications in terms of organisational development, building new competences and publishing proactive guidance that are possible outcomes from the collaboration.

Keep it simple

Start small, evaluate and change as appropriate. If one tries to consider all aspects before doing anything, it can easily be overwhelming. It is also possible to run innovation-facilitating initiatives without investing in a technical infrastructure. In fact, we recommend not investing in technical infrastructure until there is a concrete need to address. Often, the involved organisations have access to the technical resources they need to develop, train, test and validate their AI systems. And if they were to adjust their operations and systems to the tools and technologies provided by an authority that would incur costs and lead times, something that could exclude actors with a more limited budget.

When approaching something new and uncharted, it can be convenient to take a step back and try to get a full picture before acting. Our experience from both projects is that it is when the participants intertwine the theoretical analysis with hands-on action that the pain and gain become tangible. The first project, PLSF, really gained momentum when the team decided to use mock-ups for trying out the Swedish Transport Agency's guidelines for trialling smart ships. That enabled us to pinpoint some insights, such as the lack of certificates for ships under five meters. It also illustrates an important aspect of the way of working, as we reduce the overall complexity to something small(er) that still holds qualities valuable to invest. From the gained insights, it is possible to increment and iterate to build a larger analysis with more details and increased scope.

Market maturity

As a first practice is sought, the outcomes need to include a broader conversation to be iterated, predictable and adopted by third parties and thereby possibly establish a best practice. The notifications and complementary safety cases for trialling smart ships are publicly available through the authority and the experience report. The report

from PLSF and the guidelines for trialling smart ships on Swedish waters are available on the authority's webpage.[17] In this way, the innovators take responsibility for market maturity by supplying their cases, and the authorities by making their guidance publicly available. When choosing to openly share the new policies and the outcomes from trialling innovative AI, team members not only take responsibility for their actions. The parties also contribute to their solutions becoming broadly known, discussed and contradicted and possibly accepted as best practice.

Market maturity is more than reasonably safe products and services provided by private enterprises in fair competition. It also relies on authorities that understand how business opportunities are changing due to new possibilities, be they technical or regulatory, so market surveillance adapts accordingly. Authorities can gain insights into how technology and business are evolving by being involved in others' innovation, and subsequently can decide on how their application of policy is to accommodate an evolving market. But if that insight is to facilitate market maturity, the development of the application of policy needs to be shared openly so all enterprises can assess, and possibly adjust, their development.

[17] The Swedish Transport Agency, Autonom sjöfart och smarta fartyg (Swedish only), https://www.transportstyrelsen.se/sv/sjofart/autonom-sjofart-och-smarta-fartyg/

References

1. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, A Competitiveness Compass for the EU, Brussels, 29.1.2025 COM (2025) 30 final. Retrieved from: https://commission.europa.eu/document/download/10017eb1-4722-4333-add2-e0ed18105a34_en?filename=Communication_1.pdf

2. International Maritime Organisation, Autonomous shipping, https://www.imo.org/en/mediacentre/hottopics/pages/autonomous-shipping.aspx

3. Black, J. (2022). Critical Reflections on Regulation, Australian Journal of Legal Philosophy.

4. Burden, H., & Stenberg, S. (2025). An intermediate governance model for EU regulatory sandboxes. Retrieved from: https://diva-portal.org/smash/record.jsf?pid=diva2:1995583

5. Burden, H., Stenberg, S., Carlgren, L., & Sjöblom, T. (2022). Policylabb Smarta Fartyg. In Swedish, but includes an English summary. Retrieved from: https://diva-portal.org/smash/get/diva2:1707210/FULLTEXT01.pdf

6. Burden, H., Stenberg, S., Carlgren, L., & Sjöblom, T. (2023). The Swedish policy lab for maritime autonomous surface ships. Transportation Research Procedia, 72, 1840–1847. Retrieved from: https://doi.org/10.1016/j.trpro.2023.11.661

7. The Swedish Transport Agency, Transportstyrelsens riktlinjer för tester med smarta fartyg, TSS 2020-4309 (Swedish only). Retrieved from: https://www.transportstyrelsen.se/globalassets/global/sjofart/autonom-sjofart/transportstyrelsens-riktlinjer-for-tester-med-smarta-fartyg.pdf

8. Burden, H., Stenberg, S., Nilsson, E., & Petersson, C. (2025). Slutrapport Policy Lab Urban Zjöfart, with English summary. Retrieved from: Retrieved from: https://diva-portal.org/smash/get/diva2:1943322/FULLTEXT01.pdf

9. Burden, H., & Stenberg, S. (2023). Sustainable AI and Disruptive Policy – AI Regulatory Sandboxes. Retrieved from: https://ri.diva-portal.org/smash/get/diva2:1835556/FULLTEXT01.pdf

10. The Swedish Authority for Privacy Protection (2024). Disclosure of Public Records Using AI, see https://www.imy.se/globalassets/dokument/rapporter/english-summary-disclosure-of-public-records-using-ai.pdf (English summary)

11. Burden, H., & Stenberg, S. (2025). EU regulatory sandboxes - An opportunity for coordinating AI Innovation. Retrieved from: https://diva-portal.org/smash/record.jsf?pid=diva2:1995581

12. The Swedish Transport Agency, Autonom sjöfart och smarta fartyg (Swedish only), https://www.transportstyrelsen.se/sv/sjofart/autonom-sjofart-och-smarta-fartyg/

SUPERVISING GUARDRAILS FOR RESPONSIBLE INNOVATION: FROM AI INCIDENTS TO RED LINES

Author: Tereza Zoumpalova; The Future Society

Abstract

Effective AI supervision requires balancing innovation with protection through clear, enforceable guardrails. This article proposes a dual-pillar framework for supervisory authorities: incident oversight and red lines. The first pillar establishes systems for early detection of AI harms through reporting channels, whistleblower protections, and user complaints, enabling authorities to contain risks before they escalate from localized incidents to systemic crises. The second pillar enforces risk-proportionate prohibitions on AI systems and uses that pose unacceptable risks to fundamental rights, safety, and democratic values. Drawing primarily on the EU AI Act as the most advanced operational framework, while maintaining global relevance, the article demonstrates how these complementary safeguards create regulatory certainty for developers and investors while protecting public trust. It further argues that because AI risks transcend borders, effective supervision demands international collaboration through mechanisms such as mutual assistance regimes, shared incident monitoring, and coordinated enforcement. By anchoring oversight in both proactive monitoring and firm prohibitions, supervisory authorities can facilitate AI development within boundaries that protect society from its most severe risks.

Introduction

AI can deliver transformative benefits, yet without effective safeguards and supervision, it will spark systemic harms. For supervisory authorities, the challenge is ensuring that innovation does not come "at any cost" but unfolds within clear and enforceable guardrails. Far from being a brake on innovation, supervision can facilitate responsible development by providing regulatory certainty and public trust.

Approaches to AI that involve moving fast can come with quick advancements, but also significant setbacks. If safety is compromised for the quick deployment of new products and features, it is more likely that some of these will cause either unacceptable risks or function well on the whole, but with accidents happening, which can lead to significant damage and backlash. These unacceptable risks include both dangerous AI uses (such as mass surveillance, AI-generated bioweapon design assistance, or AI systems used to manipulate children) and harmful AI behaviours (where systems might engage in unauthorised self-replication, break into computer systems, or resist human control)[1]. Such risks could lead to widespread and irreversible harms, from engineered pandemics to systematic human rights violations. To maintain public trust in the products and services and protect their users, it is necessary to accept a more cautious approach to innovation, which prioritises responsibility and operating within guardrails that ensure safety.

If guardrails are clearly defined and consistently enforced, they create an environment where developers can innovate with confidence that their systems align with expected norms. Investors, in turn, gain assurance that products can scale without unexpected accidents, potential future regulatory hurdles, or reputational damage. Much of the uncertainty comes from grey zones, i.e., areas that may not yet be regulated but could later be deemed unacceptable or subject to stricter rules. At the global scale, for instance, untargeted facial recognition scraping from social media to build databases existed in a regulatory grey zone for years before multiple jurisdictions began restricting such practices. In such cases, long-term returns become difficult to predict. The problem is compounded by regulatory fragmentation, which forces companies to navigate a patchwork of national rules for the same technologies. Similarly, the EU AI Act aims to harmonise rules across Member States,[2] while establishing national supervision systems that support safe, cross-border innovation.

[1] Russell, S., Segerie, C.-R., Iliadis, N. and Zoumpalova, T. (2025). AI governance through global red lines can help prevent unacceptable risks - OECD.AI. [online] OECD AI Wonk. Available at: https://oecd.ai/en/wonk/ai-governance-through-global-red-lines-can-help-prevent-unacceptable-risks.

[2] This is something that can be seen from the wording of the EU AI Act. In its first sentence, the text lays out its purpose with clear references to the aims to support responsible innovation by facilitating the functioning of the internal market while upholding ethical values: "The purpose of this Regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, the placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union, in accordance with Union values, to promote the uptake of human centric and trustworthy artificial intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (the 'Charter'), including democracy, the rule of law and environmental protection, to protect against the harmful effects of AI systems in the Union, and to support innovation. This Regulation ensures the free movement, cross-border, of AI-based goods and services, thus preventing Member States from imposing restrictions on the development, marketing and use of AI systems, unless explicitly authorised by this Regulation

This article mostly takes a jurisdictional focus on the EU, given that the EU AI Act (AIA) is currently the most advanced effort to operationalise supervision, combining harmonised rules across Member States with national oversight mechanisms. The EU thus provides a concrete example of how incident reporting systems and red lines can be embedded in law and supervisory practice. At the same time, the analysis is of global relevance. Other jurisdictions will face similar challenges regarding how to enable innovation while enforcing limits on unacceptable risks, and they may look to the EU as a reference point. In this sense, the EU model offers an early blueprint for supervisory authorities worldwide.

While legislation establishes the rules, supervisory authorities are uniquely positioned to bring these rules to life through their legally mandated powers of enforcement and oversight. Unlike other actors who may voluntarily monitor AI systems or advocate for safety, supervisory authorities alone hold the legal mandate to require compliance, investigate incidents, demand corrective action, and impose penalties for violations. This makes them essential gatekeepers in translating regulatory frameworks into real-world protections. This article, directed at supervisory authorities themselves, argues that their effective oversight rests on two important complementary safeguards.

The first pillar is incident oversight: systems for reporting, monitoring, and investigating early warnings of harm. Supervisory authorities possess the legal authority to mandate incident reporting from AI providers, conduct formal investigations, and compel remedial measures. By catching problems before they escalate (such as through whistleblower channels, red-team findings, or user complaints), supervisory authorities can prevent small failures from becoming systemic risks.

The second pillar is red lines: enforceable prohibitions against AI uses that pose unacceptable risks, such as applications that undermine human control or enable mass manipulation. Red lines in AI governance can be defined as "specific, non-negotiable prohibitions on certain AI behaviours or AI uses that are deemed too dangerous, high-risk, or unethical to permit. These boundaries are intended to protect the survival, security, and liberty of humankind."[3] Here too, supervisory authorities play the critical enforcement role by interpreting prohibitions, identifying violations, and ensuring compliance through their sanctioning powers. Together, these guardrails strike a balance. They allow space for responsible innovation while setting clear boundaries that protect society. Anchoring oversight in both proactive monitoring and firm prohibitions gives supervisory authorities the credibility and capacity to operationalise public trust.

Taken together, the article argues that effective AI supervision hinges on two complementary pillars: incident oversight and risk-proportionate red lines. It shows how supervisory authorities can detect and contain emerging harms through robust reporting channels, whistle-

[3] Zoumpalova, T. and Iliadis, N. (2025a). Part 1: What Are Red Lines for AI and Why Are They Important? - The Future Society. [online] The Future Society. Available at: https://thefuturesociety.org/airedlines-partone.

blower protections, user-complaint systems, and crisis-response tools, while also enforcing clear prohibitions on AI behaviours and uses that pose unacceptable risks. Drawing on the EU AI Act as the most advanced operational model, but with relevance beyond Europe, the article sets out how these mechanisms can be implemented in practice, how authorities should interpret and enforce red lines as technologies evolve, and why international cooperation is essential for managing cross-border risks. In doing so, the article aims not to advocate for red lines on AI in the abstract, but to clarify how supervisory authorities can operationalise them, alongside incident oversight, to enable responsible innovation within enforceable boundaries.

Incident Reporting & Prevention: Building Early Warning and Trust

Incidents and Escalation Risks

Hazards such as unsafe design features or malicious uses can escalate quickly. Left unmanaged, small-scale incidents can snowball into sectoral emergencies or even cross-border crises. Supervisory authorities, therefore, play a critical role in ensuring that risks are captured early in the escalation pathway and contained before they destabilise essential systems or erode public trust.

The escalation pathway of AI incidents[4] describes how risks evolve in severity:

- Hazards are potential risks where harm has not yet occurred, but vulnerabilities exist.
- Incidents are contained cases where AI causes localised harm.
- Emergencies are fast-moving threats that overwhelm normal responses, often with national security implications.
- Crises are systemic, cross-border disruptions causing mass harm and destabilising societies.

This framework shows how small hazards, if unmanaged, can cascade into full-scale crises. In this context, we can see supervisory authorities being at the first line of defence, spotting AI hazards and responding to incidents. Under the EU AI Act, this responsibility is formalized through mandatory incident reporting requirements. Article 73 of the AIA requires providers of high-risk AI systems to immediately report serious incidents to the market surveillance authorities (MSAs) of the Member States where the incident occurred, without undue delay. These MSAs are then obligated to report such incidents to the European Commission and other relevant authorities, ensuring that information flows through the supervisory network. This structured reporting obligation ensures that incidents are captured systematically rather than remaining isolated events, enabling authorities

[4] Gor, G. and Iliadis, N. (2025). What Is an Artificial Intelligence Crisis and What Does It Mean to Prepare for One? - The Future Society %. [online] The Future Society. Available at: https://thefuturesociety.org/aicrisisexplainer/.

to identify patterns, coordinate responses, and prevent localized failures from escalating into wider crises.

As in many other fields, containing the risks early on is the most effective solution. For instance, when it comes to public health outbreaks, threats are best contained in the earliest stages, such as when a disease is starting to spread. Similarly, if AI hazards are spotted early on through these mandatory reporting channels, they can be addressed before starting to cause large-scale harm.

AI could lead to incidents with far-reaching consequences in the physical world. Case study 4 in UNESCO's second training set for supervisory authorities provides an example that could be used to illustrate this pathway.[5] The scenario is about Faux-Face, a technology used to create highly realistic deepfakes. At the hazard stage, the mere existence of such a system creates vulnerabilities, i.e., its potential for misuse is clear even before any harm occurs. The case study then describes an incident in which a deepfake of a well-known public figure announced a major stock market crash. This localised incident triggered panic selling, leading to significant market disruption. Left unchecked, such incidents could escalate further: rapid contagion effects in financial markets risk becoming emergencies, overwhelming normal regulatory and economic safeguards. In the worst case, cascading economic instability could reach the level of a full-scale crisis, with systemic cross-border repercussions.

To prevent this escalation, supervisory authorities must act at the earliest stages of the pathway, leveraging the powers granted to them under the AI Act. In the Faux-Face deepfake scenario described above, market surveillance authorities would need to investigate whether the system complied with transparency requirements. Article 50 of the AIA establishes comprehensive transparency obligations for AI systems that generate or manipulate image, audio, or video content, requiring that such content be marked in a machine-readable format and detectable as artificially generated or manipulated. For the Faux-Face case specifically, this would mean ensuring clear labelling of the deepfake content as machine-generated, making it immediately identifiable to viewers and platforms, thereby reducing the risk of the financial panic that occurred.

In cases of non-compliance with these transparency requirements, MSAs have the authority under Article 20 to impose corrective measures, ranging from requiring the provider to bring the system into compliance to more severe actions such as withdrawal from the market or recall of the system. By exercising these enforcement powers at the hazard or early incident stage (before a deepfake-induced market disruption can trigger broader financial contagion), authorities can contain risks and prevent incidents from spiralling into full-scale crises.

[5] Duller, Y. and Fernandez, A. (2025). Enhancing governance: use cases for supervisory authorities (Training 2). [online] UNESCO UNESDOC Digital Library. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000393209?posInSet=2&queryId=0182d2c9-6a4f-4aa9-8fa7-d87b99958685.

For supervisory authorities tasked with monitoring and preventing incidents involving systems in the European Union, understanding the definitions and classifications under the EU AI Act is crucial. Consistent treatment of the definitions can prevent both under-reporting (hence not giving some incidents sufficient attention) and over-reacting (beyond the powers of the supervisory authorities or the needs of the situation).

A "serious incident" is precisely defined in Article 3(49) of the AIA as "an incident or malfunctioning of an AI system that directly or indirectly leads to any of the following:

(a) the death of a person, or serious harm to a person's health;

(b) a serious and irreversible disruption of the management or operation of critical infrastructure.

(c) the infringement of obligations under Union law intend to protect fundamental rights;

(d) serious harm to property or the environment."

This definition is critical because, when such an incident occurs, providers of high-risk AI systems are obliged by Article 73 of the AIA to immediately report it to the market surveillance authorities of the Member States where the incident happened. The report must be made without undue delay, and in cases of widespread infringement or death, specific expedited timelines apply, such as within two days for widespread infringement/critical infrastructure disruption or ten days for a death, once a causal link or reasonable likelihood is established or suspected. Following such a report, the provider is required to conduct necessary investigations, including a risk assessment and take corrective actions, in cooperation with competent authorities.

Beyond these specific serious incidents, the AI Act employs a risk-based approach to categorise AI systems broadly, recognising their potential to pose risks to safety, fundamental rights, and health. This classification dictates the level of regulatory scrutiny and compliance obligations. AI systems are generally categorised into minimal, limited, high-risk, and unacceptable risk.

An important part of the AIA in this context is about high-risk AI systems, which, while not prohibited, are subject to stringent requirements. An AI system is classified as high-risk if it is intended to be used as a safety component of a product or is itself a product covered by specific Union harmonisation legislation listed in Annex I of the AIA and requires a third-party conformity assessment. Additionally, AI systems explicitly listed in Annex III AIA are considered high-risk. Examples of high-risk systems under Annex III can include those for remote biometric identification, systems evaluating creditworthiness, AI used in recruitment or selection of persons, medical AI applications, AI for dispatching emergency services, and systems influencing election out-

comes. Supervisory authorities play a vital role in ensuring responsible AI[6] development and deployment by accurately categorising AI systems and enforcing the associated requirements and reporting obligations.

Reporting and Spotting Incidents

To effectively spot incidents as defined under the AI Act (and importantly, to identify potential hazards before they materialise into incidents), secure channels of reporting are essential. Here, "hazards" refer to the earlier stage in the escalation pathway: potential risks where vulnerabilities exist but harm has not yet occurred, whereas "incidents" under Article 3(49) are events where harm has already materialised (such as death, serious health harm, or fundamental rights infringements). Catching hazards early, before they cross into incidents, is critical for prevention.

Supervisory authorities play a crucial role in establishing secure channels for both incident and hazard reporting under the EU AI Act. By "secure channels," we mean both the technical infrastructure (such as dedicated reporting platforms) and the procedural frameworks (including confidentiality protections, clear reporting pathways, and defined responsibilities) that enable safe and effective information flow. The EU already has precedent for such systems: under NIS2 (the Network and Information Security Directive), entities must report significant cybersecurity incidents through designated platforms that ensure confidential handling and coordinated response. Similarly, the Cyber Resilience Act (CRA) establishes incident notification mechanisms for products with digital elements. These existing frameworks demonstrate how secure, platform-based reporting combined with clear procedures can enable timely information sharing while protecting sensitive data and maintaining trust among reporting parties. For AI supervision, similar infrastructure is needed. While the Act mandates providers of high-risk AI systems to report "serious incidents" to market surveillance authorities without undue delay, and general-purpose AI model providers to report serious incidents and corrective measures to the AI Office and national competent authorities, this mandatory provider reporting should be supplemented by channels accessible to others who can see the risks, including whistleblowers, users, and civil society, who may detect hazards or incidents that providers have not yet identified or reported.

Whistleblowers are often the earliest and most reliable source of warnings about unsafe AI practices.[7] Insiders can spot weak security, misleading claims, or rushed deployments long before such issues become visible to regulators or the public. Yet speaking up carries high personal risk, from retaliation to career loss, and protection re-

[6] "Responsible AI" is understood differently across stakeholders, from ethical principles and voluntary commitments to binding legal obligations, and it can be a vague term. In this article, it denotes AI development within enforceable regulatory guardrails rather than aspirational principles: systems that enable early hazard detection through incident reporting and respect prohibitions on unacceptable risks.
[7] Ryan, F. and Zoumpalova, T. (2025). Why Whistleblowers Are Critical for AI Governance - The Future Society. [online] The Future Society. Available at: https://thefuturesociety.org/ai-whistleblowers.

mains uneven. While the EU Whistleblower Protection Directive provides a baseline within Europe, enforcement is patchy, and most jurisdictions worldwide lack comparable safeguards. To counter this, supervisory authorities should not treat whistleblowing as separate from incident reporting, but embed secure and protected channels directly into their systems. This integration implicitly necessitates systems for whistleblower protection, often involving secure platforms that may include options for anonymity to safeguard reporting individuals. In a Statement from the Chairs and Vice Chairs of the Safety and Security Chapter in the European Union's General-Purpose AI Code of Practice, a recommendation mentioned a proposal[8] to establish a dedicated reporting channel for AI-related whistleblower disclosures. Such an addition could be of significant help in enabling people to safely report their safety concerns.

All authorities involved in the application of the AI Act are required to respect the confidentiality of information and data obtained, particularly intellectual property, confidential business information, and trade secrets. Furthermore, these authorities must implement adequate and effective cybersecurity measures to protect the security and confidentiality of the information and data obtained. Doing so both taps into insider knowledge and ensures early-stage hazards are surfaced before escalating into full-blown crises.

As an example, case study 9 in UNESCO's training set for supervisory authorities describes MedAIPro, an AI system which has been found to misdiagnose some conditions.[9] In this example case, former anonymous engineers from the developing company came forward, raising information about how validation testing was rushed due to tight deadlines. Such an example shows how whistleblowers can bring and strengthen claims about safety concerns. To keep AI safe, we need to make sure that people with such information can safely come forward.

Just as importantly, it is necessary to pay attention to AI hazards spotted by users of the systems. Citizen complaints serve as crucial oversight tools, with the AI Act explicitly allowing any natural or legal person to lodge a complaint with the relevant national market surveillance authority (MSA) if they have grounds to consider an infringement of the Act (article 85). To maximise their effectiveness, authorities should establish accessible, user-friendly complaint procedures, like web portals and help desks, and actively promote the public's right to complain, thereby creating a broader detection network for potential AI breaches. This could be similar to how the rights to request compensation for delayed flights or trains are promoted in Europe.[10] Travellers can easily come across such information at airports

[8] In a statement from the chairs and vice-chairs leading the drafting of the Safety & Security Chapter of the Code of Practice, the authors include a recommendation for the AI Office to establish dedicated reporting channels for whistleblowers who wish to submit AI-related disclosures.
[9] Duller, Y. and Fernandez, A. (2025). Enhancing governance: use cases for supervisory authorities (Training 2). [online] UNESCO UNESDOC Digital Library. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000393209?posInSet=2&queryId=0182d2c9-6a4f-4aa9-8fa7-d87b99958685.
[10] For situations like train or flight delays, EU citizens can easily learn about their rights by clicking through official EU websites (e.g. here for rail and here for air travel) and based on the situation that they select, they see what compensation they have rights to.

and train stations, so they are more likely to be aware of their rights and of the procedures for filing complaints and requesting compensation.

Non-EU authorities should similarly define clear reportable thresholds, such as "serious harm or rights breaches," to ensure both industry and citizens can effectively report issues into the oversight system. Even if they do not yet have any legal obligations to track such reports, if they have sufficient capacity to do so, they could take significant steps forward by setting up such reporting systems to be aware of potential hazards and incidents, helping them protect their citizens.

Even with strong preventative safeguards, serious AI incidents will still occur, requiring supervisory authorities to act swiftly. Market surveillance authorities already have the powers to investigate, mandate corrective actions, or withdraw non-compliant systems, but these need to be operationalised through crisis playbooks. Such plans should define escalation triggers, coordination protocols across agencies, and transparent communication strategies that maintain public trust without fuelling panic. Learning can be found in high-risk industries, ranging from aviation and medicine. Systematic logging of anomalies and proactive crisis planning can turn early warnings into visibility of AI risks and their effective containment. But visibility alone cannot prevent the deployment of fundamentally unacceptable systems. For this, supervisory authorities need the complementary pillar: risk-proportionate enforcement of red lines that prohibit the most dangerous practices.

Red Lines: Enforcing Risk-Proportionate Prohibitions

Defining and Identifying Red Lines

Incident reporting helps detect risks in motion, but some systems pose such fundamental dangers that they must never be deployed in the first place. In the EU, this means enforcing the AI Act's prohibitions on unacceptable-risk systems, interpreting these provisions in emerging contexts, and identifying potential gaps. Internationally, jurisdictions at different stages of developing AI-specific legislation can leverage existing legal frameworks, such as data protection laws, consumer protection regulations, anti-discrimination statutes, and human rights protections, many of which already prohibit practices that the AI Act explicitly addresses in the AI context. For example, manipulative practices targeting vulnerable groups or discriminatory automated decision-making may already be illegal under existing law, even without AI-specific provisions. Supervisory authorities can thus play a vital role in identifying how AI applications intersect with these established legal protections, supporting the adaptation and strengthening of regulatory frameworks to address AI-specific risks,

and contributing to international dialogue on defining clear red lines for AI uses and behaviours that pose unacceptable risks.res from escalating into wider crises.

Such red lines can fall under two categories:[11]

1. AI behaviours: "Limits on certain behaviours that AI systems should not exhibit (e.g., developers must prove that their systems will not exhibit self-replication or improvement without human control, even if the systems are technically capable of doing so)."
2. AI uses: "Limits on how humans can use the AI system (e.g., a prohibition on using AI to manipulate or surveil children, even if technologies enabling doing so exist)."

Such red lines have been gaining significant attention from a broad range of individuals and organisations calling for them. The broader public has expressed concern over AI and a desire for some dangerous AI systems or uses to be prohibited in a public consultation of over 10,000 citizens in the preparations of the 2025 AI Action Summit in France[12]. In the same report, over 200 expert organisations were consulted, and they also frequently expressed the urgency of establishing prohibitions around unacceptable risks. This echoes a consensus statement of global AI scientists from March 2024, when they gathered in Beijing and jointly called for AI red lines, focusing especially on catastrophic risks stemming from AI behaviours.[13]

Red lines in the EU

Despite the interest in drawing red lines on unacceptable risks, such prohibitions are only starting to emerge.[14] However, in the EU, some already exist in practice and are therefore of practical interest to supervisory authorities.

The EU AI Act lays out prohibitions on AI systems considered unacceptable. Based on Article 5 of the AIA, this includes:

- "Exploitation of vulnerabilities of persons, manipulation and use of subliminal techniques;
- Social scoring for public and private purposes;
- Individual predictive policing based solely on profiling people;
- Untargeted scraping of the internet or CCTV for facial images to build up or expand databases;

[11] Zoumpalova, T. and Iliadis, N. (2025a). Part 1: What Are Red Lines for AI and Why Are They Important? - The Future Society. [online] The Future Society. Available at: https://thefuturesociety.org/airedlines-partone.
[12] The Future Society (2024). Citizens and Experts Unite: Final Report on Global Consultations for France's 2025 AI Action Summit - The Future Society. [online] The Future Society. Available at: https://thefuturesociety.org/aiactionsummitconsultationreport/.
[13] IDAIS (2024). IDAIS-Beijing - International Dialogues on AI Safety. [online] International Dialogues on AI Safety. Available at: https://idais.ai/dialogue/idais-beijing/.
[14] Zoumpalova, T. and Iliadis, N. (2025b). Part 2: Are There Red Lines for AI in Practice Already? - The Future Society. [online] The Future Society. Available at: https://thefuturesociety.org/airedlines-parttwo.

- Emotion recognition in the workplace and education institutions, unless for medical or safety reasons (i.e. monitoring the tiredness levels of a pilot);
- Biometric categorisation of natural persons to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs or sexual orientation. Labelling or filtering of datasets and categorising data in the field of law enforcement will still be possible;
- Real-time remote biometric identification in publicly accessible spaces by law enforcement, subject to narrow exceptions (under a special status with requirements to aim for this prohibition to protect fundamental rights, avoiding misuse and overreach while enabling law enforcement)."

These prohibitions entered into force on the 2nd of February, 2025.

Various safety-oriented red lines are not yet implemented in AI governance but already show signs of international support, such as those from the Beijing Consensus Statement on Red Lines in Artificial Intelligence (IDAIS, 2024). These include red lines around autonomous replication or improvement, power seeking, assisting weapon development, cyberattacks and deception. Such proposed red lines currently go beyond existing EU prohibitions, but are examples of potential future needs that some experts are calling for.

The Role of Supervisory Authorities

Supervisory authorities in the EU play a critical role not only in enforcing the AI Act's explicit prohibitions but also in interpreting these red lines as new technologies and use cases emerge. While Article 5 lists specific banned practices (such as manipulative behavioural techniques, social scoring, or untargeted facial recognition), authorities must assess whether novel applications fall under these prohibitions when risks materialise in new forms, such as advanced emotional recognition or AI-enabled mass surveillance.

This requires combining legal interpretation with technical expertise, supported by instruments like audits, impact assessments, and AI registers, to detect when emerging capabilities cross into "unacceptable risk." In this way, supervisory authorities act as the first line of defence, ensuring that red lines remain robust and adaptable to evolving technological contexts rather than becoming static or outdated.

Supervisory authorities in the EU have a layered toolkit to identify and respond to potential red lines under the AI Act. Their oversight starts with preventive instruments. Providers of high-risk AI systems must maintain extensive technical documentation, logs, and record-keeping obligations, all of which can be inspected by authorities. The Act also establishes conformity assessments and the EU database of high-risk AI systems (the AI register), which allow authorities to monitor what is on the market and spot potential violations. In addition,

authorities can carry out audits and inspections if they suspect that a system may be in breach, making it possible to detect when a practice falls under the Act's explicit prohibitions in Article 5.

Once a possible violation is identified, authorities move into the enforcement phase. If they determine that an AI system constitutes a prohibited practice, Article 20 empowers them to impose corrective measures, including ordering the provider or distributor to bring the system into compliance, withdraw it from the market, or recall it altogether.

Where providers fail to act, supervisory authorities themselves can take measures to restrict or prohibit the system's availability. In cases of serious or systemic breaches, Chapter XII provides for penalties, with fines reaching up to C35 million or 7% of global annual turnover for violations of Article 5 prohibitions. This strong sanctioning power underpins the credibility of red lines by ensuring that non-compliance carries significant costs.

Supervision is reinforced by the Union safeguard procedure in Article 81, which enables rapid escalation: if one national market surveillance authority bans a prohibited system and the European Commission finds it justified, the Commission can extend this restriction across the single market.

Examples for Supervisory Authorities

Another one of the UNESCO training cases illustrates how supervisory authorities might respond when a potential red line under the EU AI Act is at stake. In the BigBank example, the company sought to deploy a system scanning public space 50 meters before customers entered a branch, raising concerns about real-time remote biometric identification[15]. Article 5(1)(h) prohibits such practices for law enforcement except under narrow conditions, while Article 5(1)(e) bans creating or expanding facial recognition databases through untargeted image scraping. A supervisory authority would investigate whether the system relied on such prohibited practices using audits, documentation checks, and AI registers. If confirmed, it could order withdrawal or recall under Article 20, and in severe cases, apply Article 83 penalties.

A contrasting case is RemoteExam, where an AI system monitors students during exams and flags “suspicious behavior” such as looking down or out of frame.[16] If the system infers emotions, such as deception or anxiety, it could fall under the prohibition on emotion recognition in education (Article 5(1)(f)). Here, authorities would need

[15] Fernandez, A. (2024). Enhancing governance: use cases for supervisory authorities (Training 1). [online] UNESCO UNESDOC Digital Library. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000391680?posInSet=1&queryId=0182d2c9-6a4f-4aa9-8fa7-d87b99958685.
[16] Fernandez, A. (2024). Enhancing governance: use cases for supervisory authorities (Training 1). [online] UNESCO UNESDOC Digital Library. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000391680?posInSet=1&queryId=0182d2c9-6a4f-4aa9-8fa7-d87b99958685.

deeper technical audits and testing to determine whether the system effectively crosses into a red line before corrective action is warranted.

This contrast shows that clear-cut cases of systems violating red lines allow swift enforcement, while borderline ones demand more interpretation and evidence-gathering. To manage both, supervisory authorities must combine technical expertise with consistent guidance across Member States, ensuring that red lines operate as enforceable standards rather than prohibitions on paper.

Red Lines Internationally

Beyond the EU's legally enforceable red lines, a much wider ecosystem of international frameworks is beginning to shape expectations for responsible AI. Many of these take the form of ethical guidelines and soft-law instruments, such as the UNESCO Recommendation on the Ethics of Artificial Intelligence, which has been adopted by 194 member states. These frameworks matter not only within the EU but also for countries without binding AI legislation, offering immediate pathways to identify unacceptable risks and build safeguards with the mandates they do have. By setting ethical standards, encouraging transparency, and flagging high-risk applications like social scoring or mass surveillance, such initiatives help institutions act early to prevent harmful practices. In doing so, they drive responsible innovation globally, ensuring that AI development not only reacts to binding prohibitions but also aligns with shared international norms and societal values.

By enforcing risk-based prohibitions wherever they exist[17], supervisory authorities are already building up the toolkit of instruments needed for potential future international red lines regimes. Momentum is building towards defining unacceptable risks and exploring how clear, verifiable and enforceable red lines could be set[18]. By leading the way by adhering to existing ethical guidelines and enforcing existing prohibitions, we are moving closer to operationalising a system of red lines that could put guardrails in place to stop the most severe risks. This way, innovation can be bounded within the context of responsible systems that are proven not to contribute to potential incidents or create unacceptable risks.

[17] While the article focuses on providing the EU AI Act's prohibitions as an example of legally binding red lines, they already exist in other jurisdictions too. An overview of existing red lines for AI can be found in a summary by The Future Society (Zoumpalova & Iliadis, 2025b).

[18] An example of this momentum is the Global Call for AI Red Lines, a campaign launched at the 80th UN General Assembly by CeSIA, The Future Society and UC Berkeley's Center for Human-Compatible AI, calling on governments to agree by the end of 2026 on clear, verifiable, enforceable prohibitions on AI uses deemed universally unacceptable. It has been signed by over 300 prominent individuals, including 10 former heads of state or ministers (e.g., Juan Manuel Santos, Mary Robinson), 15 Nobel laureates (e.g., Maria Ressa, Joseph Stiglitz, Geoffrey Hinton) and Turing Award winners (e.g., Yoshua Bengio, Andrew Yao); and AI pioneers (e.g., Ian Goodfellow, Wojciech Zaremba). It has also been backed by over 90 organisations.

International Collaboration: From Local Oversight to Global Red Lines

Need for Cross-Border Supervision

AI systems are often developed in one jurisdiction, deployed in another, and used globally. National oversight alone cannot prevent cascading harms. Supervisory authorities, therefore, need international coordination mechanisms to share data, align on red lines, coordinate on monitoring incidents, build preparedness frameworks, and act collectively when crises loom.

The International AI Safety Report 2025 warns that flaws in widely deployed general-purpose systems can cascade rapidly: a single model update or release may simultaneously affect millions of users across sectors, with impacts that are sudden and potentially irreversible.[19]

Because many of these systems are deployed globally, their risks transcend borders. An incident in one country can quickly trigger harm elsewhere, much like financial shocks that spill into global recessions or pathogens that cause pandemics. Examples include a malfunctioning AI model disrupting global supply chains, automated cyberattacks propagating through interconnected networks, or disinformation systems destabilising multiple democracies at once. These risks are interlinked, as failures in one sector or country can set off cascading effects in others, which underscores the urgent need for coordinated international safeguards.

Avenues for International Collaboration

For supervisory authorities, international collaboration is not optional but essential. The EU AI Act already embeds a principle of mutual assistance. If a high-risk or prohibited system is flagged in one Member State, authorities elsewhere are obliged to share evidence, documentation, or even coordinate inspections. The EU AI Office adds another layer, serving as a coordination hub to align enforcement practice, issue guidance, and ensure consistent interpretations of red lines across the Union. For national supervisors, this means their work will not remain isolated, as investigation and enforcement can be shared through a wider cooperative framework.

Beyond Europe, a growing set of international fora provides platforms for cooperation. The OECD, the UN, and the International Network of AI Safety Institutes all facilitate technical exchanges. UNESCO's Global Forum on the Ethics of AI has recently gone further by launching the Global Network of AI Supervisory Authorities (GNAIS), creating a dedicated space for supervisory authorities worldwide to collaborate and exchange lessons. For supervisory authorities, this is

[19] Y. Bengio, et al. (DSIT 2025/001, 2025). International AI Safety Report. [online] DSIT. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025.

an opportunity to both strengthen their own capacity and contribute to global alignment. Effective AI regulation does not rest solely on well-drafted laws, but on well-resourced, networked authorities that can identify cross-border risks and coordinate timely responses.

There are instructive lessons from other risk domains where international coordination has successfully prevented localized incidents from escalating into systemic crises. The WHO's International Health Regulations (IHR) framework, formalized after the SARS outbreak in the early 2000s, mandates that member states report disease outbreaks within 24 hours through centralized surveillance systems. This enables rapid identification of emerging patterns across multiple jurisdictions that individual countries might miss in isolation. The framework's success rests on binding reporting obligations, standardized data formats, and clear escalation procedures, which are principles that could apply equally to AI supervision.

For AI supervisory authorities, the WHO model offers a relevant template: shared monitoring dashboards with standardized incident reporting, enabling early detection when similar AI failures emerge across different Member States, combined with regular stress tests simulating coordinated responses to international AI incidents. Such joint exercises improve preparedness and help authorities link scattered incidents that may signal systemic flaws requiring coordinated intervention, much as pandemic surveillance detects emerging threats before they spread globally.

By collaborating across borders, supervisory authorities can extend their reach and ensure that their local enforcement contributes to a safer global AI ecosystem.

Conclusion

Supervisory authorities can stand at the intersection of innovation and protection by ensuring that AI develops within guardrails that are both credible and enforceable. The dual pillars of oversight presented here, incident reporting and the enforcement of risk-proportionate red lines, provide some of the foundations for this role. By capturing hazards early through secure reporting channels, audits, and whistleblower protections, authorities prevent localised harms from escalating into systemic crises. By enforcing prohibitions on unacceptable uses, they create certainty and signal to innovators where the boundaries lie, while ensuring that the public is not endangered by high-risk AI uses and behaviours. Together, these mechanisms transform supervision from a perceived brake on innovation into a framework that sustains trust and enables responsible progress.

However, AI risks do not stop at national borders, so isolated supervision cannot suffice. Authorities must therefore embrace international collaboration as part of their core mandate. The EU AI Act's mutual-as-

sistance regime, the coordinating role of the AI Office, and global fora such as UNESCO's GNAIS provide the scaffolding for this cooperation.

Learning from other fields (from nuclear early warning to pandemic surveillance), supervisors can build joint preparedness plans, incident simulations, and shared monitoring systems to anticipate and contain transnational harms.Ultimately, strong supervision provides developers with the certainty that their products can scale responsibly, investors with confidence that risks are managed, and citizens with assurance that their rights and safety are protected. By embedding transparency, proportionality, and international cooperation into their oversight, supervisory authorities ensure that AI remains a tool for progress rather than disruption.

References

Duller, Y. and Fernandez, A. (2025). Enhancing governance: use cases for supervisory authorities (Training 2). [online] UNESCO UNESDOC Digital Library. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000393209?posInSet=2&queryId=0182d2c9-6a4f-4aa9-8fa7-d87b99958685.

Fernandez, A. (2024). Enhancing governance: use cases for supervisory authorities (Training 1). [online] UNESCO UNESDOC Digital Library. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000391680?posInSet=1&queryId=0182d2c9-6a4f-4aa9-8fa7-d87b99958685.

Gor, G. and Iliadis, N. (2025). What Is an Artificial Intelligence Crisis and What Does It Mean to Prepare for One? - The Future Society %. [online] The Future Society. Available at: https://thefuturesociety.org/aicrisisexplainer/.

IDAIS (2024). IDAIS-Beijing - International Dialogues on AI Safety. [online] International Dialogues on AI Safety. Available at: https://idais.ai/dialogue/idais-beijing/.

Russell, S., Segerie, C.-R., Iliadis, N. and Zoumpalova, T. (2025). AI governance through global red lines can help prevent unacceptable risks - OECD.AI. [online] OECD AI Wonk. Available at: https://oecd.ai/en/wonk/ai-governance-through-global-red-lines-can-help-prevent-unacceptable-risks.

Ryan, F. and Zoumpalova, T. (2025). Why Whistleblowers Are Critical for AI Governance - The Future Society. [online] The Future Society. Available at: https://thefuturesociety.org/ai-whistleblowers.

The Future Society (2024). Citizens and Experts Unite: Final Report on Global Consultations for France's 2025 AI Action Summit - The Future Society. [online] The Future Society. Available at: https://thefuturesociety.org/aiactionsummitconsultationreport/.

Y. Bengio, et al. (DSIT 2025/001, 2025). International AI Safety Report. [online] DSIT. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025.

Zoumpalova, T. and Iliadis, N. (2025a). Part 1: What Are Red Lines for AI and Why Are They Important? - The Future Society. [online] The Future Society. Available at: https://thefuturesociety.org/airedlines-partone.

Zoumpalova, T. and Iliadis, N. (2025b). Part 2: Are There Red Lines for AI in Practice Already? - The Future Society. [online] The Future Society. Available at: https://thefuturesociety.org/airedlines-parttwo.

SECTION IV.

CROSS-SECTORAL COORDINATION MECHANISMS

As supervisory authorities begin to develop interpretative capacities, experiment through sandboxes, and embrace more innovative forms of oversight, a structural question naturally follows: how can these efforts be connected across institutions and across borders? AI supervision cannot stay confined within sectoral silos or national boundaries. Because AI systems interact with multiple legal regimes and circulate across jurisdictions, coordination becomes an essential pillar of effective supervision.

The discussions at the Expert Roundtable underscored this underlying reality. They highlighted, above all, that coordination must be understood as operating at two levels. The first is interinstitutional coordination within a country: the ability of data protection authorities, competition regulators, consumer protection bodies, and sectoral supervisors to align their approaches when overseeing AI systems. This requires more than occasional information exchange. It calls for shared interpretative practices, common reference points, and institutional arrangements that help authorities navigate overlapping mandates without duplication or contradiction. Formal structures, such as thematic networks or coordination councils, can provide this stability; agile, informal working groups can complement it by enabling rapid engagement when new issues arise.

This theme was the focus of the fourth session of the Expert Roundtable, which examined how coordination mechanisms can translate the promise of AI regulation into operational practice. The guiding question of "How do we design cross-sectoral coordination mechanisms?" captured a growing concern among regulators: that fragmented oversight structures risk undermining both the consistency of enforcement and the confidence of those subject to it. Participants agreed that coordination is not a procedural luxury but a condition for credible supervision.

The second level is coordination across countries. No single jurisdiction can build complete supervisory capacity in isolation, nor can any authority detect or understand emerging risks purely within its own borders. Cross-country cooperation allows regulators to observe how similar systems behave in different contexts, learn from each other's enforcement experiences, and identify patterns or vulnerabilities that might otherwise remain hidden. It also prevents a fragmented landscape in which operators encounter divergent expectations or seek out jurisdictions with the weakest oversight. In this sense, cooperation is not only an efficiency gain; it is a safeguard for coherence, credibility, and fairness in global AI markets.

What stood out from the Roundtable was not a call for one specific model of coordination, but a recognition that coordination must be intentional and multi-layered. AI supervision demands mechanisms that combine legal clarity with institutional trust: arrangements that give authorities a stable basis for joint interpretation while allowing enough flexibility to adapt to technological change. Clarity for supervisors, in turn, creates clarity for operators and deployers. When authorities articulate expectations coherently, within sectors, across sectors, and across borders, they reduce uncertainty, support compliance, and enable more predictable and responsible innovation.

The contributions in this section build on these insights. They discuss the institutional conditions that make cooperation effective, the challenges authorities face when navigating across sectors, and the kinds of structures, formal and informal, that can sustain cross-sectoral and cross-country collaboration over time. Together, they offer a vision of cooperative supervision that reflects the complexity of AI itself: distributed, adaptive, and sustained by trust among institutions.

FROM SILOS TO SYNERGY: CYBERSECURITY LESSONS IN CROSS-SECTORAL COORDINATION

Author: Carlos Moreira Antunes; Portuguese National Cybersecurity Centre

Abstract

A regular critique of emerging technologies and cybersecurity governance is that supervision remains overly compartmentalised, exhibiting excessive verticalization at the expense of cross-sectoral coordination. While siloing assets or activities can sometimes be justified by considerations of quality, quantity, or incompatibility, such isolation often undermines elements whose value depends on connectivity, interoperability, and co-evolution within broader systems.

On this matter, this paper would like to advance the argument that supervisory fragmentation stems primarily from prevailing methodologies and organisational cultures, rather than from inherent technical constraints. In public administrations, the prevailing institutional reflex when facing novel challenges – such as artificial intelligence (AI) or cybersecurity – is to decompose them into narrowly defined, specialised domains. The absence – at times almost deliberate – of a return to the whole, enabling reintegration based on structural or functional linkages, generates disruption and entrenches thematic silos. This results in a form of "laboratory regulation" paradigm that perpetuates systemic fragmentation.

Within the European Union (EU), governance cultures shaped by legality and subsidiarity often prioritise debates on institutional responsibility ("who does it") over substantive deliberation on needs and appropriateness ("what must be done"). To address these limitations, this paper proposes adopting a lifecycle approach as a guiding regulatory principle. By framing phenomena across their entire developmental continuum, governance can move beyond a static, siloed "warehouse" model toward a dynamic, holistic "watershed" perspective that manages interdependent flows.

Drawing on the experience of the Portuguese National Cybersecurity Centre (CNCS), this study, which actively fosters collaboration across regulators, industry, and academia, demonstrates the effectiveness of cross-sectoral regulatory methodologies that prioritise the regulatory object over fragmented institutional perspectives.[1] The paper concludes by proposing a model of collaborative supervision, applicable nationally and internationally, grounded in the principle that regulatory authorities within common markets should operate cooperatively rather than competitively.

[1] The Portuguese National Cybersecurity Authority, a participant in the Working Group for Competent Authorities on AI and chair of its cybersecurity workstream.

EU Member States are encouraged to strengthen cross-sectoral coordination in emerging technologies, particularly AI and cybersecurity, where regulation has often been reactive and fragmented. A more holistic and integrated approach is needed. This paper approaches this crucial topic from three perspectives: first, a situational awareness that enables it to identify the real demands that must be addressed; second, the specific issue of law and subsidiarity in this context; and third, a proposal of a lifecycle approach methodology.

Situational Awareness

Being mindful of the characteristics of emerging technology markets and of the demands inherent to the associated regulatory activity constitutes a kind of situational awareness that leads to more effective, forward-looking decision-making. This paper argues that, in fact, it is preferable to start from an empirical observation of the technology itself, rather than from an immediate focus on existing regulatory instruments. Indeed, one of the most frequently cited challenges concerning technology is the slower evolution of law, creating what some authors refer to as the "pacing problem."[2] However, this should not lead to the conclusion that regulation is, per se, objectively negative.

It becomes clear that a defining feature of emerging technology markets is their vocation for universality – a sort of democratisation of access to technology – since they appear to be aimed at reaching as many users as possible. As a brief historical glance at computers and networks shows, the creation and public availability of the TCP and IP protocols undeniably opened the doors to an expanding web of connections, constrained only by the capacity of the infrastructures themselves.[3] The first incidents and risks gave rise to cybersecurity, highlighting the need to protect users – especially those less able to detect or respond to threats.

The universalisation of technology has blurred the distinction between the physical and digital worlds, making digital interaction the default reality. In cyberspace, solutions and products from a single source are simultaneously accessible to users across multiple contexts, underscoring the "ubiquity" of digital technologies such as AI.[4] In cyberspace, even if there are boundaries – within whatever limits that can meaningfully be called boundaries – the reality is that solutions, products, or technologies are made available from a single creator to a multitude of users, across dozens of cyberspaces, simultaneously. Curiously, unlike clothing or cars – for which brand and country of origin matter greatly – in emerging technologies, networks, or systems, what matters most is whether they are accessible. There is a certain anonymity of the deployer or provider, with little concern for who they are.

[2] A Reuel and TA Undheim, "Generative AI and Adaptive Governance" (2024) arXiv <https://doi.org/10.48550/arXiv.2406.04554> accessed 18 August 2025, 3.
[3] D Van Puyvelde and AF Brantly, Cybersecurity, Politics, Governance and Conflict in Cyberspace (Polity 2025) 12.
[4] OJ Erdélyi and J Goldsmith, "Regulating Artificial Intelligence: Proposal for a Global Solution" (2020) Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society <https://arxiv.org/abs/2005.11072v1> accessed 18 August 2025, 1.

New technologies naturally foster connections – mimicking their very origins – creating single points of access and reinforcing interdependence across critical sectors such as energy, health, finance, and transport. While this connectivity brings clear benefits, it also amplifies vulnerabilities and threats. Indeed, as infrastructure systems become increasingly interlinked across sectors, the failure of a single component can trigger cascading effects across others, amplifying disruption.[5] Modern critical infrastructures are not isolated but deeply interdependent, meaning that disruptions – whether technical failures or cyberattacks – can spread rapidly across sectors, multiplying the impact and jeopardising resilience. This paper, therefore, deals with "amplifiers of fragilities," as the field has, unfortunately, experienced on several occasions, observing the scaling effects of an attack or vulnerability on interdependent or interconnected infrastructures, of which the 2017 WannaCry ransomware attack is a prime example, reaching 150 countries in a single day.[6] [7]

Such risks are exponential in the case of AI, due to its capacity for uninterrupted operation, the scale of users it reaches, the rapid dissemination of data or links, but also the overconfidence of users in its outputs and potential, bias, its tendency to always provide an answer, and its short-term effectiveness – without overlooking a certain fascination with the technology itself.[8] [9] Emerging technologies introduce additional vulnerabilities that warrant careful consideration. These arise not only from their inherent operational characteristics – such as opacity, discriminatory tendencies, and "black box" architectures – but also from the spectrum of potential threats exploitable by malicious actors, such as data poisoning, model poisoning, adversarial examples, model evasion, as well as confidentiality breaches, systemic model failures, membership inference, and prompt injection.[10] Beyond their technical manifestations, these vulnerabilities intersect with broader governance challenges, as their detection, mitigation, and accountability mechanisms often lag the pace of technological deployment. As the NIST framework highlights, "[u]nderstanding and managing the risks of AI systems will help to enhance trustworthiness, and in turn, cultivate public trust."[11]

Added to these factors are the unintended uses and unforeseen effects of technology – arising either from negligence or from a lack

[5] GAO, Technology Assessment: Cybersecurity for Critical Infrastructure Protection (GAO-04-321, 2004) <https://www.gao.gov/products/gao-04-321> accessed 20 August 2025<https://www.gao.gov/products/gao-04-321> accessed 20 August 2025;
[6] M Suleyman, A Próxima Vaga (Clube do Autor 2023) 189.
[7] Europol, Internet Organised Crime Threat Assessment (IOCTA) 2018 (Publications Office of the EU 2018) <https://www.europol.europa.eu/internet-organised-crime-threat-assessment-2018> accessed 21 August 2025; Symantec, WannaCry: Ransomware Attacks Show Strong Links to Lazarus Group (Symantec Security Response 2018) https://www.symantec.com/blogs/threat-intelligence/wannacry-ransomware-attack. accessed 19 August 2025
[8] EM Bender, T Gebru, A McMillan-Major and S Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (2021) Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency 610.
[9] L Floridi and M Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences" (2020) 30 Minds and Machines <https://doi.org/10.1007/s11023-020-09548-1> accessed 20 August 2025, 681.
[10] J Dupuy, "Legal Transparency in AI Finance: Facing the Accountability Dilemma in Digital Decision-Making" Reuters (1 March 2024) https://www.reuters.com/legal/transactional/legal-transparency-ai-finance-facing-accountability-dilemma-digital-decision-2024-03-01/
[11] NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1, National Institute of Standards and Technology 2023) <https://www.nist.gov/itl/ai-risk-management-framework> accessed 18 August 2025, 1.

of understanding of its full potential. As Deibert and Rohozinski aptly observe, "[t]echnical innovations may be designed for specific purposes but often end up having wildly different social uses and effects than those intended by their creators."[12] This observation underscores the necessity for adaptive regulatory frameworks capable of responding to emergent behaviours and impacts beyond the scope originally anticipated by developers. A further feature of emerging technologies is their free availability, with few legal restrictions or licensing requirements, enabling rapid market entry and innovation.

While this openness represents an extraordinary advantage, it also exponentially accelerates the volume of new applications and realities made accessible to the public. Such dynamism places considerable pressure on regulators, who are compelled to keep pace with continuous technological advancements and shifting trends, lest regulatory oversight fall irreparably behind.

Technological universalism, therefore, clashes with the rigid compartmentalisation of regulatory and supervisory bodies. Instead of ensuring clarity, this fragmentation often distracts innovators and diverts investment and research to other regions, undermining regulatory effectiveness. In the field of cybersecurity, for instance, regulation encompasses the Cybersecurity Act, NIS2, DORA, CRA, and the AI Act, alongside all specific national legislations. This is precisely the point of interest and the foundation for the (lack of) success of regulation in general. Perhaps due to a certain explanatory simplification, there is a tendency to emphasise the idea of over-regulation, which may indeed hold some truth in terms of the size and segmentation of regulatory texts, but there is often a failure to grasp that the success of laws lies in the institutional and human components that enforce them, as demonstrated by the EU AI Act, which, in several recitals and articles, has required the strengthening of technical, financial, and human resources to make the adopted regulation effective.[13] [14]

Regulatory challenges in emerging technologies stem from multiple factors, such as weak communication channels, cultural and trust deficits, misaligned incentives, fragmented regulatory regimes, structural barriers, limited strategic intelligence, and recurring preventable errors. Recognising these issues is key to designing strategies for effective cross-sectoral coordination.

This regulatory approach, grounded in a silo or "warehouse" logic, decomposes the landscape into narrowly defined and specialised domains, making them easier to manage but neglecting mechanisms of structural or functional reintegration. The resulting paradigm of "laboratory regulation" reinforces systemic fragmentation and under-

[12] RJ Deibert and R Rohozinski, "Liberation vs. Control: The Future of Cyberspace" (2010) 21(4) Journal of Democracy 43, <https://www.journalofdemocracy.org/articles/liberation-vs-control-the-future-of-cyberspace/> accessed 20 August 2025, 44.
[13] F Brito Bastos, "Regulatory Administrative Law and the Digital Era: The Case for a Pluralistic Research Agenda" (2025) 32(3) Maastricht Journal of European and Comparative Law, <https://doi.org/10.1177/1023263X251366573> accessed 18 August 2025, 239.
[14] C Novelli and others "A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities" (arXiv, 2024) <https://doi.org/10.48550/arXiv.2407.10369> accessed 22 August 2025, 12.

mines overall regulatory coherence.[15] The purpose of regulation is the realisation of the public interest within a given sector. This interest must not be confused with the institutional interests or conveniences of regulatory authorities, which are themselves shaped by legal, human, and organisational dynamics. Equally, regulatory activity should avoid any form of regulatory capture, where private or sectoral interests overshadow the public interest. Regulators must therefore refrain from asserting authority for their own sake or prioritising operators' concerns at the expense of broader societal goals.

Any regulation addressing markets and sectors characterised by rapid transformation, universality, and a growing proliferation of economic actors and creators must seek to respond to these structuring elements within its very architecture. Situational awareness is essential for comprehensive decision-making and effective regulation. Accordingly, this paper may recommend to the authorities that:

(i) They should encourage the establishment of intersectoral observatories tasked with producing information and policy papers regularly, thereby supporting planning in relation to supervisory actions, as well as normative and regulatory adjustments.

(ii) They should support or create mechanisms, through multidisciplinary structures involving researchers and innovators, for the testing and monitoring of products, systems, and networks after their market deployment, to diagnose vulnerabilities preventively.

That said, this paper would like to address two aspects highly relevant for understanding future regulatory activity: the Principle of Legality and the Principle of Subsidiarity.

Reframing the Principles of Legality and Subsidiarity

As Brito Bastos notes, by recalling that "[t]oday, it is beyond dispute that administrative law, across all Member States, has undergone an extensive process of Europeanization."[16] Within this movement, two principles emerge as essential: legality and subsidiarity. However, prevailing interpretations of this principle are frequently embedded within an overly positivistic background, relying heavily on traditional notions of centralised authority. This approach constrains the capacity for adaptive governance in dynamic and interdependent policy environments. As the OECD highlights, modern governance requires "strategic, evidence-based and innovative policies to strengthen public governance," which suggests that a rigid reading of legality risks undermining regulatory responsiveness.[17]

[15] An expression we use to describe a regulatory framework that still appears experimental, learning by doing, rather than following a stable methodology or producing a definitive outcome.
[16] Brito Bastos (n 12) 242.
[17] OECD, Recommendation of the Council on Regulatory Policy and Governance (OECD Publishing 2012), <https://www.oecd.org/en/publications/recommendation-of-the-council-on-regulatory-policy-and-governance_9789264209022-en.html> accessed 22 August 2025, 3.

Regulation and regulatory authorities are a well-established feature worldwide. Critical sectors, given their impact on people and enterprises, require sustained oversight to ensure quality, resilience, and stability. Public authorities employ various tools to this end, from supervision and monitoring to enforcement and sanctioning powers. Naturally, such powers and prerogatives are exercised by individuals or institutions with formal responsibility for defining and implementing them. It is precisely here that challenges arise. Sectoral regulation may take place within a purely national framework, but in many cases – especially when it concerns critical infrastructures or essential services in domains that transcend both physical borders and cyberspace – the regulatory dimension necessarily becomes transnational. In areas of political or economic integration, such as the European Union, regulatory requirements, standards, and binding legal instruments are defined at the supranational level.

However, these alone do not exhaust the regulatory action required. Once regulations, directives, standards, or even international conventions (such as the UNESCO Recommendation on the Ethics of Artificial Intelligence) are adopted, States remain responsible for implementing such provisions within their own sectors and services, primarily through their administrative structures.[18]

Two immediate conclusions:

(i) It is the administrative authorities of individual states – established and organised in accordance with their sovereign prerogatives – that apply these provisions and translate them into concrete rules and procedures; and

(ii) Supranational rules and standards, while binding, inevitably require processes of harmonisation and adaptation within domestic legal systems before they can be effectively operationalised.

The first signs of fragmentation arise from the principle of legality, a cornerstone of administrative law and democratic safeguards. It requires that administrative entities act only within the law, a principle codified in national constitutions and reflected in international instruments through provisions such as legal reservation, procedural guarantees, and oversight mechanisms. Even in cases where regulatory agencies are endowed with a degree of discretion, their authority is still exercised strictly within the legal framework. However, interpretations of the principle of legality that remain tied to a rigid form of legal positivism can transform the law into a constraint or source of inertia. This is especially problematic in sectors marked by constant change, where excessive formalism may hinder timely regulatory responses and the capacity to adapt to evolving risks. As noted by Mark Hodgins, “legislative processes broadly employ procedural safe-

[18] UNESCO, Recommendation on the Ethics of Artificial Intelligence (adopted 23 November 2021) https://unesdoc.unesco.org/ark:/48223/pf0000381137.

guards [...] but the safeguards delay regulatory enactment as well", illustrating a phenomenon that has been termed "regulatory sclerosis."[19]

Indeed, it is important to reaffirm that the principle of legality itself is not under question; it remains fundamental to effective public administration and to the proper functioning of a democratic rule of law. What is at stake, rather, is the way in which this principle should be interpreted and applied in highly technical and security-sensitive domains such as emerging technologies and cybersecurity.

Another crucial element lies in the application of the principle of subsidiarity. This principle aims to ensure greater effectiveness of administrative action by bringing decision-makers closer to those affected by their decisions. It is enshrined, for example, in Article 5(3) of the Treaty on European Union.[20] Subsidiarity offers clear advantages: it allows more immediate responses from the closest administrative and decision-making levels, thereby avoiding excessive centralisation. Indeed, emerging technologies and cybersecurity have benefited from this scaling effect, as they have increasingly been analysed and regulated in broader markets than the national one, notably at the European level.

However, the interaction between legality and subsidiarity can lead to regulatory behaviours that are more reactive than proactive. National regulators may become dependent on supranational structures, which can result in two main tendencies: (i) limiting themselves to reacting to legal stimuli emanating from international institutions, without fundamentally rethinking the internal legal framework; or (ii) interpreting European norms through an overly national lens, thereby producing regulatory fragmentation.[21]

Consequently, these factors tend to shift the regulatory debate from "what needs to be done" towards "who should do it." In many instances, this debate emerges too late or becomes dependent on methodologies and processes already defined at the supranational level, thereby reducing the room for manoeuvre of national administrative entities. Frequently, such entities arrive only at the very final stage of discussions in which the practical application of European regulations is being shaped – a timing that undermines the effectiveness of regulation in critical and fast-evolving sectors such as artificial intelligence and cybersecurity.

As is widely understood, these principles and concepts were designed and developed primarily for "analogic" administrations and for contexts in which sovereignty was clearly defined. It is therefore unsur-

[19] M Hodgins, "The Perils of Cybersecurity Regulation" (2024) Review of Austrian Economics <https://doi.org/10.1007/s11138-024-00660-4> accessed 18 August 2025, 3.2.
[20] Consolidated Version of the Treaty on European Union [2012] OJ C326/13, art 5(3).
[21] M Lodge and K Wegrich, Managing Regulation: Regulatory Analysis, Politics and Policy (Palgrave Macmillan 2012) <https://www.researchgate.net/publication/286879237_Managing_Regulation_Regulatory_Analysis_Politics_and_Policy> accessed 20 August 2025, 18

prising that such principles may enter crisis when applied to digital domains, where a plurality of stakeholders operate without a clear territorial link and where sovereignty itself is contested, as in cyberspace.[22]

A shift is needed towards a service- and responsibility-based paradigm, where legality and subsidiarity act not only as limits but as enablers of cooperative, responsive, and citizen-centred governance. This requires rethinking regulatory mechanisms beyond strict legal dictates or traditional administrative structures. As stated by the World Economic Forum, "[a]ddressing these evolving threats demands not only advanced technological solutions but also cross-sector collaboration and knowledge-sharing."[23] This highlights the necessity of complementing legal principles with cooperative mechanisms across jurisdictions and sectors.

Member States are encouraged to consider the following approaches:

(i) Strengthening collaboration with academia to develop multidisciplinary training and dynamic supervisory tools – such as the use of advanced AI models – that integrate normative sources and support adaptive, networked governance.

(ii) Fostering legal creativity and safe testing environments to prevent the consolidation of inadequate legal solutions and to promote their continuous improvement through stakeholder collaboration.

Towards an Object-Centred Model of Regulatory Supervision

More important than knowing whom to hold accountable if things go wrong is understanding who must be at the table to ensure that things go right, and even before asking who should sit at the table, first asking why a table should be convened at all.

Concerning the essential characteristics of emerging technologies and, by extension, cybersecurity, this allows the understanding of the choice that many legal systems have made in favour of a risk-based model of regulation or supervision. Such a model proves to be comparatively faster in delivering responses than regulatory or supervisory approaches centred primarily on economic actors or on the prior approval of technologies and products.

This orientation reflects broader international debates on the governance of emerging technologies, particularly the tension between ex ante approval models – which require prior certification of technologies or products before they enter the market – and ex post or risk-based models, which focus on identifying, monitoring, and mitigating risks as they materialise.

Nevertheless, it must be acknowledged that the idea of a risk-ba-

[22] Puyvelde and Brantly (n 2), 50.
[23] World Economic Forum, Global Risks Report 2025 (WEF 2025) <https://www.weforum.org/publications/global-cybersecurity-outlook-2025/> accessed 20 August 2025, 42.

sed approach remains largely conceptual and generic. In practice, the regulatory regimes applied to AI and cybersecurity, though both nominally grounded in risk, diverge significantly in substance. AI oversight is predominantly centred on legal risks or rights-based concerns – even if only formally, because dependent on "lists" – whereas risks in the field of cybersecurity are primarily technical or procedural.[24] Thus, the very notion of risk takes on different meanings and requires distinct regulatory approaches.

In the case of AI, regulation is essentially tied to documentary and quality obligations in the development and deployment of high-risk AI systems, with virtually no barriers to their release on the market. Cybersecurity, by contrast, follows what Mei and Sag describe as the logic of the "immunological other" – a preventive and proactive stance aimed at anticipating and responding to cyberthreats that extend far beyond intentional attacks (secure by design).[25] Indeed, cybersecurity has evolved under considerable standardisation efforts and the imposition of mandatory security measures calibrated by risk levels, reflecting a systematic effort at anticipation. This is consistent with the very definition of cyberthreat under the Cybersecurity Act.[26] Such a definition encompasses much more than deliberate actions by malicious actors.

What both fields do share, however, is a reliance – sometimes criticised as excessive – on generic and abstract concepts, as well as a fragmented and complex regulatory framework, which is particularly problematic when the regulatory target itself is moving and constantly evolving.[27]

Such a circumstance undermines trust as a factor of innovation. Trust is a prerequisite for innovation, investment, and global technological progress. It is built structurally through clear and consistent rules, and procedurally through transparent and replicable methodologies that ensure fairness and predictability. International organisations have repeatedly stressed that building trustworthy AI governance requires a human-rights-based approach.[28] Equally significant is the human dimension of trust, which requires the presence of competent, accountable, and credible interlocutors at each stage of technological development and investment. In this sense, trust operates not only as a condition for effective cross-sectoral coordination but also as a catalyst for transnational cooperation, enabling regulatory convergence, facilitating the flow of investments, and fostering sus-

[24] Arnoud Engelfriet, The Annotated AI Act: Article-by-Article Analysis of European AI Legislation (ICTRecht 2024) 16.
[25] Yiyang Mei and Matthew Sag, The Illusory Normativity of Rights-Based AI Regulation (arXiv, March 2025) <https://doi.org/10.48550/arXiv.2503.05784> accessed 20 August 2025, 28.
[26] 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 [2019] OJ L151/15, art 2(8).
[27] PAE Davis and others, EU Cybersecurity Regulation in the Quantum Age in Christopher Markou and Matthias Leese (eds), Quantum Technology Governance: Law, Policy and Ethics in the Quantum Era (Springer Nature Singapore 2025, advance online publication) <https://doi.org/10.2139/ssrn.5383838> accessed 19 August 2025, 4.2.
[28] UNESCO, Recommendation on the Ethics of Artificial Intelligence (2021) <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence> accessed 21 August 2025.

tainable innovation. Without such multi-layered trust, both the stability of global investment ecosystems and the resilience of technological infrastructures are inevitably compromised.

This paper would therefore like to propose a model of cross-sectoral coordination that considers the multiple aspects previously described. In fact, excessive detail in the drafting of measures, without considering sector-specific features, risks becoming ineffective. What is needed, rather, is a bird's-eye view of the supervision of emerging technologies, recognising that they are framed within a broader spectrum – a digital space – of which they are just one component.

Moreover, a network-like functioning, comparable to that of a watershed, requires acknowledging that any obstruction of water flows or the building of dams prevents the renewal and circulation of matter. Fragmentation occurs precisely because the supervisory model is based on "water reservoirs" or "warehouses" without permanent flow – that is, water is passed from one container to another, with part being lost along the way and impurities accumulating in the process. Thus, authorities frequently accumulate responsibilities and knowledge (silos) but transfer them to their counterparts or to interdependent sectoral authorities only when deemed strictly necessary, and then immediately disconnect – warehouse, or perhaps more accurately, a form of purely bureaucratic behaviour. This represents a transition from a silo-based model to one of synergy.

One of the key lessons learned from cybersecurity is the importance of network-based functioning and of information-sharing mechanisms that ensure the right stakeholders are brought to the table. Working in a crisis room, or, more broadly, in a multidisciplinary team, is essential, since every issue always has multiple dimensions and therefore requires multiple interventions.

By aligning stakeholders, defining scope, setting goals, and integrating both technical and non-technical perspectives, organisations create networked and synergistic governance structures. This approach not only ensures compliance but also enables more adaptive, resilient, and holistic supervision. Such an approach is grounded in the lifecycle principle: the focus is not limited to a specific aspect or detail of the technology or system but instead addresses the whole. In this model, it is the authorities who go to the technology, just as doctors visit patients, rather than forcing the "patient" to move from one isolated authority to another.

In this same vein, some authors, such as Erdélyi and Goldsmith, have proposed a global solution model based on the creation or restructuring of an existing Intergovernmental Organisation that would ensure the coordination of public policy development at the national level.[29] This paper believes it is necessary to go further and recognise the need for a true coordination model, with a moderating instituti-

[29] Erdélyi and Goldsmith (n 3), 12.

on capable of convening, redirecting, and connecting structures within a network, avoiding vertical hierarchies. Both AI and cybersecurity cannot be regulated through vertical silos or a single dimension, but rather across multiple layers.

At the same time, clarity of roles within the network and the establishment of regular multidisciplinary or inter-institutional actions, with mandatory intermediate approval requirements, create forms of permanent cooperation. It must be acknowledged that, when it comes to emerging technologies and cybersecurity, competent authorities are more effective when focusing on prevention and the adoption of best practices from the outset, rather than relying solely on sanctioning mechanisms or excessively rigid requirements. This amounts to a friendly and proactive form of supervision.

This paper considers it appropriate to adopt a model in which an entity or a competence centre, acting as coordinator, regularly brings together all stakeholders involved in the reporting process linked to the technology lifecycle (some models already published for the AI Act reflect this logic, such as the models of the Netherlands and Cyprus):

Coordinator Entity

- Acts as the central convener of all stakeholders.
- Regularly organises joint meetings and reporting sessions.
- Builds bridges across institutions and sectors.

Functions

- Propose cooperation teams.
- Design shared deliverables on specific topics.
- Collect and distribute information at national and international levels
- Ensure constant development, adaptation, and updating of practices.

Key Tool: “Regulatory Card”

- Allows full traceability of the regulatory process across the technology lifecycle.
- Testing phase.
- Market release.
- Updates and patches.
- Incidents and failures.
- Use cases and objectives.
- Monitoring and inspection.
- Regulatory dialogue.

Advantages

- Rapid identification of legal and technical competencies.
- Avoids duplication of procedures, resources, and acquisitions.
- Supports a genuine national strategy.
- Promotes cross-sectoral governance and prevents fragmentation or disruption.

Naturally, such a model cannot be confined to public entities alone, which is why the creation of listening and sharing forums with regulated actors and economic stakeholders is of utmost importance. This model may produce two very practical immediate effects: (i) the preparation of proposals for legal and institutional reform, thereby adapting the legal and organisational framework to evolving contexts and market realities while avoiding "functional fragmentation"; and (ii) the approximation of governance models among States, particularly in areas of legal, cultural, and economic convergence, through stronger connections between authorities.[30]

These mechanisms will, in the short term, allow fragmented economic spaces to evolve from vertical or piecemeal legislation (as often happens in Europe) towards a model of codification. This does not preclude, however, the need to review the multiple governance models that have been created in supranational regulation on a case-by-case basis and without a holistic vision – as can be seen in the clear distinction between the governance model for cybersecurity regulation and that of AI.

But shall there once again be inaction, waiting for the law to change, thereby confirming the legal positivism already discussed? No! Action can be taken proactively – a term preferable to informal, which is inaccurate – through the initiative of entities and authorities themselves, both at the national and international levels. It is worth sharing some of the lessons learned, particularly within the domain of cybersecurity. One of the key drivers of fragmentation lies in institutional and organisational culture – the regulatory stance adopted toward legislation and, above all, toward the regulated. This can be reshaped within a lifecycle approach model by embracing two guiding principles: proactivity and collaboration. Proactivity requires authorities to engage early and systematically with international and national counterparts to address cross-sectoral regulatory challenges through joint forums. Collaboration calls for the open sharing of methodologies and solutions, reflecting the very logic of the open-source principle. This logic reflects the idea that once one authority takes the lead, others are encouraged to follow. The first mover creates momentum, setting a precedent that fosters wider adoption of good practices across the network.

Some examples of the positive effects of adopting this methodology already exist and can be studied. One example is the Working

[30] François Delerue, Cyber Operations and International Law (CUP 2020) 19.

Group for Competent Authorities on AI (WGCAAI), which proactively brings together authorities that wish to assemble to discuss issues related to AI, its supervision, and to establish methodologies, share knowledge and tools, raise questions for debate, and organise small specialised groups based on specific topics, which then share their results with the plenary.

This paper can also share some of the Portuguese experience, particularly from the work of the National Cybersecurity Centre (CNCS), which has paid special attention to the supervision of emerging technologies. The CNCS has not only recognised the importance of proactively participating in international and national forums for multidisciplinary coordination, such as the NIS Cooperation Group (formal) and WGCAAI (proactive), but has also established mechanisms of cooperation and capacity-building that reduce the risks of fragmentation regarding the cybersecurity regulatory framework. This paper highlights the following:

C-Network – Network of Cybersecurity Competence Centres

- Supports the development of national cybersecurity capabilities.
- Guides organisations, particularly SMEs, towards maturity and resilience via seven regional centres.

Guidelines and Frameworks

- Developed through participatory methodologies integrating market requirements and international best practices.

Cybersecurity Observatory

- Multidisciplinary systematisation of information across society, economy, public policy, ethics, law, risks, and innovation.

Cybersecurity Communities

- Sectorial and thematic entity networks aimed at building trust and fostering information sharing.
- Provide intelligence, events, and communication platforms.

Legal Framework and High Council for Cyberspace Security

- Ensures political-strategic coordination among policy makers, sectoral authorities, law enforcement, judiciary and defence.

C-Days Conference

- Annual partnership with academia to share best practices and strengthen national cyberspace resilience.

C-Academy Platform

- Advanced cybersecurity training for public administration and th private sector, in collaboration with higher education institutions.

In conclusion, the effective implementation of a coordinated and cross-sectoral model requires a holistic approach that overcomes the compartmentalisation fostered by vertical regulation and fragmented international acts. Advancing a culture of collaboration – and rejecting individualism – can accelerate this process, as regulation entails no competition but the collective pursuit of public interests. The paper can end with one important question: if technology flows like a river, why regulate it from a warehouse?

References

Books

A Engelfriet, The Annotated AI Act: Article-by-Article Analysis of European AI Legislation (ICTRecht 2024)

DV Puyvelde and AF Brantly, Cybersecurity, Politics, Governance and Conflict in Cyberspace (Polity 2025)

F Delerue, Cyber Operations and International Law (CUP 2020)
M Lodge and K Wegrich, Managing Regulation: Regulatory Analysis, Politics and Policy (Palgrave Macmillan 2012) <https://www.researchgate.net/publication/286879237_Managing_Regulation_Regulatory_Analysis_Politics_and_Policy> accessed 20 August 2025

M Suleyman, A Próxima Vaga (Clube do Autor 2023)

Journal Articles

A Reuel and TA Undheim, "Generative AI and Adaptive Governance" (arXiv, 2024) <https://doi.org/10.48550/arXiv.2406.04554> accessed 18 August 2025

C Novelli and others, "A Robust Governance for the AI Act: AI Office, AI Board, Scientific Panel, and National Authorities" (arXiv, 2024) <https://doi.org/10.48550/arXiv.2407.10369> accessed 22 August 2025

EM Bender and others, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" (2021) Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency <https://doi.org/10.1145/3442188.3445922> accessed 19 August 2025
F Brito Bastos, "Regulatory Administrative Law and the Digital Era: The Case for a Pluralistic Research Agenda" (2025) 32(3) Maastricht Journal of European and Comparative Law <https://doi.org/10.1177/1023263X251366573> accessed 18 August 2025

L. Floridi and M Chiriatti, "GPT-3: Its Nature, Scope, Limits, and Consequences" (2020) 30 Minds and Machines <https://doi.org/10.1007/s11023-020-09548-1> accessed 20 August 2025

MW Hodgins, "The Perils of Cybersecurity Regulation" (2024) Review of Austrian Economics <https://doi.org/10.1007/s11138-024-00660-4> accessed 18 August 2025

OJ Erdélyi and J Goldsmith, "Regulating Artificial Intelligence: Proposal for a Global Solution" (2020) Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society <https://arxiv.org/abs/2005.11072v1> accessed 18 August 2025

PAE Davis and others, “EU Cybersecurity Regulation in the Quantum Age” in C Markou and M Leese (eds), Quantum Technology Governance: Law, Policy and Ethics in the Quantum Era (Springer Nature 2025, advance online publication) <https://doi.org/10.2139/ssrn.5383838> accessed 19 August 2025

RJ Deibert and R Rohozinski, “Liberation vs Control: The Future of Cyberspace” (2010) 21(4) Journal of Democracy <https://www.journalofdemocracy.org/articles/liberation-vs-control-the-future-of-cyberspace/> accessed 20 August 2025

Y Mei and M Sag, “The Illusory Normativity of Rights-Based AI Regulation” (arXiv, 2025) <https://doi.org/10.48550/arXiv.2503.05784> accessed 20 August 2025

Official Reports and Institutional Documents

Europol, Internet Organised Crime Threat Assessment (IOCTA) 2018 (Publications Office of the EU 2018) <https://www.europol.europa.eu/internet-organised-crime-threat-assessment-2018> accessed 21 August 2025

GAO, Technology Assessment: Cybersecurity for Critical Infrastructure Protection (GAO-04-321, 2004) <https://www.gao.gov/products/gao-04-321> accessed 20 August 2025

NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1, 2023) <https://www.nist.gov/itl/ai-risk-management-framework> accessed 18 August 2025

OECD, Recommendation of the Council on Regulatory Policy and Governance (OECD Legal Instruments 2012) <https://www.oecd.org/en/publications/recommendation-of-the-council-on-regulatory-policy-and-governance_9789264209022-en.html> accessed 22 August 2025

Symantec, WannaCry: Ransomware Attacks Show Strong Links to Lazarus Group (Symantec Security Response 2018) <https://www.symantec.com/blogs/threat-intelligence/wannacry-ransomware-attack> accessed 19 August 2025

UNESCO, Recommendation on the Ethics of Artificial Intelligence (UNESCO 2021) <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence> accessed 21 August 2025

World Economic Forum, Global Cybersecurity Outlook 2025 (2025) <https://www.weforum.org/publications/global-cybersecurity-outlook-2025/> accessed 20 August 2025

News and Media

J Dupuy, “Legal Transparency in AI Finance: Facing the Accountability Dilemma in Digital Decision-Making” Reuters (1 March 2024) <https://www.reuters.com/legal/transactional/legal-transparency-ai-finance-facing-accountability-dilemma-digital-decision-2024-03-01/> accessed 17 August 2025

SUPERVISING AI THROUGH COOPERATION AMONG COMPETENT AUTHORITIES

Author: Marc Rotenberg; Center for AI and Digital Policy

Abstract

Supervisory authorities responsible for the oversight of artificial intelligence (AI) are emerging at a time when advanced systems are already embedded in critical public and private functions. Countries are experimenting with different institutional models: some have created new, dedicated AI agencies; others rely on existing regulators (data protection authorities, consumer protection bodies, sector regulators or human rights commissions) that are expanding their mandates to cover automated decision-making. Despite this institutional diversity, many of the supervisory tasks are converging: assessing high-risk systems, reviewing documentation and testing, monitoring for discriminatory or harmful outcomes, and ensuring meaningful human oversight. Because AI systems are routinely deployed across borders and updated at speed, these authorities cannot work effectively in isolation. Cooperation is necessary to make sense of complex technical systems, avoid duplication of effort, and ensure that the lessons learned in one jurisdiction inform practice elsewhere. A cooperative approach reduces the risk of regulatory fragmentation, helps address capacity gaps, and promotes consistent oversight across jurisdictions.

The chapter situates cooperation within the UNESCO Supervising AI by Competent Authorities initiative, which aims to support supervisory authorities through capacity-building, exchange of experience, and the development of good practices. It examines existing cooperation models in adjacent fields such as data protection, financial regulation, consumer protection, and telecommunications, and considers how these can inform cooperation in AI supervision. It then describes a set of practical mechanisms that can support cooperation in practice: structured information-sharing, mapping of supervisory responsibilities, annual surveys, good practice development, engagement with external experts, and regular meetings. It highlights the need for inclusive cooperation that reflects the diversity of supervisory authorities worldwide, including those with limited resources and those in the Global South.

The chapter stresses that cooperation must be inclusive. Supervisory authorities in smaller jurisdictions or with limited technical resources face particular challenges when supervising complex AI systems developed and deployed by large multinational organisations. Cooperation allows these authorities to draw on the experience of peers, access shared tools and guidance, and participate meaningfully in shaping the norms and standards that will govern AI. At the same

time, cooperation must respect institutional independence and national legal frameworks. The challenge is to design cooperative mechanisms that support supervisory authorities without undermining their autonomy. A Global Network for AI Supervisory Authorities, supported by UNESCO and working in close dialogue with existing regional networks, can help meet this challenge by providing a service for coordination, exchange and mutual support.

Ultimately, effective cooperation enables supervisory authorities to fulfill their core purpose: ensuring that AI systems are developed and used in ways that are safe, fair and consistent with fundamental rights. It allows authorities to respond more quickly to emerging risks, to build a shared understanding of good practices, and to promote accountability in an area where technological change is rapid and unevenly distributed. A cooperative supervisory ecosystem is therefore not only more efficient; it is also more likely to uphold democratic values, protect vulnerable groups, and contribute to a more responsible and equitable AI ecosystem.

Introduction

Artificial intelligence is transforming public administration, economic life and social interaction. Governments deploy AI systems to manage public benefits, support law enforcement, allocate health resources and inform policy decisions. Private organisations use AI for credit scoring, recruitment, targeted advertising and content moderation. In many cases, these systems operate at scale, influence access to essential services and shape public discourse. The resulting benefits and risks are distributed unevenly, with particular consequences for marginalised and vulnerable communities. This raises questions of accountability: who is responsible when AI systems fail, when they discriminate, or when they erode human autonomy and dignity? Supervisory authorities sit at the centre of this emerging accountability framework. They are tasked with ensuring that AI systems comply with legal standards, respect human rights and remain transparent, accountable, and aligned with fundamental rights.

Early assessments of national readiness for AI governance increasingly emphasise the importance of independent supervision. Institutional indicators, such as the presence of an autonomous data protection authority, a consumer protection agency or a human rights commission with powers that extend to automated decision-making, are strong predictors of a jurisdiction's ability to manage AI-related risks. These authorities bring experience in handling complaints, conducting investigations, issuing guidance and coordinating with other public bodies. They also understand how to both promote innovation and protect fundamental rights. As studies of AI and democratic values have shown, jurisdictions that lack such supervisory institutions often struggle to address AI-related harms in a timely and effective manner. They may adopt ambitious AI strategies but lack the tools

to ensure that innovation occurs within the boundaries of law and public trust. Recent reports from the Organisation for Economic Co-operation and Development (OECD) on governing with artificial intelligence and on the implementation of the European Union's Coordinated Plan on AI reach similar conclusions, highlighting the need for clear institutional mandates, adequate resources, and coherent national AI governance models.[1]

Supervisory authorities face a set of challenges that are remarkably similar across jurisdictions. They must evaluate complex technical systems that may be opaque even to their developers. They must understand how AI systems are trained, validated, deployed and updated, and how different components, such as data sources, model architectures and user interfaces, interact to produce outcomes. They must assess whether safeguards, including human oversight and post-deployment monitoring, are effective in practice. They must also respond to complaints from individuals who may have been adversely affected by AI systems, even when those individuals cannot easily understand the technical details. In addition, supervisory authorities must navigate institutional constraints, including limited budgets, staffing limitations and overlapping mandates with other regulators. These challenges are amplified by the cross-border nature of many AI systems: a system deployed by a provider in one country can affect individuals in many others, and risks identified by one authority may be relevant to many others.

Cooperation, therefore, becomes a structural necessity. In supervising data protection, financial markets, consumer rights and telecommunications, authorities have long confronted similar cross-border challenges and have responded by developing cooperation mechanisms. Data protection authorities collaborate through the Global Privacy Assembly, formal agencies such as the European Data Protection Board, and regional networks including the Asia Pacific Privacy Authorities and the Network of African Data Protection Authorities. Financial regulators participate in supervisory colleges that oversee cross-border financial institutions. Consumer protection authorities coordinate through the International Consumer Protection and Enforcement Network. Telecommunications regulators cooperate through regional organisations and via the International Telecommunication Union (ITU).[2] This logic also underpins more recent international initiatives, including the G7 Hiroshima AI Process Comprehensive Policy Framework and its Guiding Principles and Code of Conduct for advanced AI systems, which explicitly call for strong regulatory and supervisory capacity and for closer cooperation among competent authorities.[3]

As AI becomes more central to public life and more complex in technical design, the need for cooperation only increases. No single

[1] OECD, 'Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions' (2025); OECD, 'Progress in Implementing the European Union Coordinated Plan on Artificial Intelligence' (2025).
[2] UNESCO, Recommendation on the Ethics of Artificial Intelligence (adopted 23 November 2021)
[3] G7, 'Hiroshima AI Process Comprehensive Policy Framework, including Guiding Principles and Code of Conduct for Organizations Developing Advanced AI Systems' (2023).

supervisory authority can be expected to have all the expertise necessary to understand and evaluate every AI system, particularly those at the cutting edge of research and development. Cooperation allows authorities to share knowledge, pool resources, and coordinate responses to cross-border issues. It also helps promote consistent interpretations of legal and ethical principles, reducing the risk of regulatory arbitrage in which providers exploit differences between jurisdictions to avoid scrutiny. Importantly, cooperation does not require uniformity. Authorities can maintain their independence while reflecting local legal and cultural contexts, while still learning from one another and aligning their approaches where appropriate. In this way, cooperation strengthens oversight without requiring harmonised laws.

This chapter explores the foundations for cooperation among AI supervisory authorities, drawing on existing practice in related regulatory fields and on emerging experience with AI supervision. It describes the different institutional models that are taking shape, examines how cooperation has functioned in other domains, and identifies practical mechanisms that can support AI supervision in practice. It also considers how cooperation can address asymmetries in resources and capacity, ensuring that authorities in smaller or less-resourced jurisdictions can participate in and benefit from global oversight efforts. Finally, it situates these developments within the broader objectives of UNESCO's capacity-building initiatives and related international efforts, including those of the OECD, the EU, and the G7, which together point toward a more coordinated and accountable global framework for AI governance.

Cooperation Among AI Supervisory Authorities

Supervisory authorities responsible for AI oversight vary substantially in their formal mandates, institutional histories and resourcing. Some are newly established agencies with a specific focus on AI. Spain's Agency for the Supervision of Artificial Intelligence, for example, has been created as a dedicated national authority with powers to supervise AI systems across sectors, working alongside specialist regulators in banking, insurance and electoral processes. Other authorities emerge from existing regulatory bodies, such as data protection authorities, consumer protection agencies or competition commissions. These authorities often expand their remit to include AI when automated decision-making and algorithmic profiling become central to their traditional domains. In still other cases, sector-specific regulators, for example, in health, transport or finance, assume AI-specific responsibilities within their respective sectors. At the same time, coordination bodies ensure coherence across the broader governance landscape.

Despite these differences, supervisory authorities share a common set of tasks. They interpret legal and ethical standards in the context of AI systems, assess risks associated with high-risk applications, evaluate

documentation and testing, and oversee post-deployment monitoring. They receive and investigate complaints, issue guidance to providers and deployers, and coordinate with courts, parliaments and other public bodies. Experience from data protection offers one of the clearest models for how authorities can carry out these tasks while cooperating across borders. Data protection authorities have long coordinated their work when individuals' personal data crosses borders. They share information about enforcement actions, cooperate on investigations and develop joint positions on emerging issues such as cross-border transfers, profiling and automated decision-making. These practices emerged because the underlying technologies – networked information systems and global data transfers – transcended national boundaries. AI supervision is now reaching a similar point.

The Global Privacy Assembly (GPA) provides a complementary model. The Assembly brings together data protection and privacy commissioners from around the world to exchange experience, adopt resolutions and coordinate on shared concerns. Over time, the GPA has established working groups on specific topics, such as technology and AI, and has adopted resolutions that articulate common expectations for controllers and processors. These instruments are not binding, but they carry moral and political weight and often influence national law and practice. The Assembly also functions as a forum for mutual capacity-building: authorities with more experience or resources can share tools and methodologies with peers who are still in the early stages of developing their regulatory frameworks. This dynamic is particularly relevant for AI supervision, where expertise is scarce, and the demand for guidance is high.

The European Data Protection Board was established by the General Data Protection Regulation, coordinating enforcement actions across the European Union by national data protection agencies and issuing formal opinions to guide interpretations of the GDPR. In the Asia-Pacific region, the Asia Pacific Privacy Authorities (APPA) forum has played a similar role. It provides a venue for authorities from countries with very different legal systems to discuss shared challenges, such as data breaches, cross-border transfers and online advertising. APPA has facilitated cooperative work on enforcement sweeps and the development of good practices, including in areas that overlap with AI, such as biometric identification and online tracking. In Africa, the Network of African Data Protection Authorities (NADPA) brings together emerging authorities from across the continent to share experience on issues such as digital identity systems, cross-border data flows and algorithmic decision-making in public services. These networks demonstrate that cooperation can flourish even when institutional capacity is uneven, provided that there is a shared commitment to mutual support and learning.

Similar patterns appear in other regulatory fields. Financial regulators oversee banks and other institutions that operate in multiple jurisdictions and can pose systemic risks if mismanaged. To address this, re-

gulators have established supervisory colleges – structured forums in which home and host authorities exchange information about a particular financial institution, conduct joint assessments and coordinate supervisory measures. These colleges rely on shared templates, regular meetings and agreed protocols for information-sharing, all of which could be adapted for AI supervision. Consumer protection authorities coordinate through the International Consumer Protection and Enforcement Network, which organises global "sweeps" of websites and online services to identify unfair or deceptive practices. Telecommunications regulators cooperate through regional organisations and through global bodies such as the ITU, allowing them to exchange information on safety standards and harmonise approaches to spectrum management and interoperability across borders, even when national regulations differ.

Across these domains, the underlying lesson is clear: cooperation is not an optional extra, but a necessary response to the cross-border nature of modern technologies and markets. Authorities that cooperate can share the burden of complex investigations, avoid duplicating effort, and adopt more consistent approaches to emerging issues. They also gain legitimacy by demonstrating that they are part of a broader community working to uphold common standards. Cooperation does not eliminate differences in law, culture or institutional design, but it provides a framework through which those differences can be managed constructively. It enables authorities to identify common ground while respecting local context, and it fosters a culture of mutual support that is essential when confronting powerful actors and complex technical systems.

AI supervision mirrors many of the complexities observed in these other domains. AI systems often operate across borders, use data from multiple sources and evolve through retraining and updates. Supervisory authorities must therefore consider not only the immediate context in which a system is deployed, but also its broader technical and organisational ecosystem. Cooperative mechanisms allow authorities to share insights into specific systems, such as widely used foundation models or common services for automated decision-making, and to coordinate their expectations for documentation, testing and oversight. When authorities in different countries align their supervisory practices, providers and deployers receive clearer signals about acceptable conduct, and individuals benefit from more consistent protection of their rights.

Cooperation is also essential for maintaining public trust. Confidence in AI governance depends not only on the existence of formal rules, but also on the perception that those rules are being applied consistently and effectively. When authorities act in isolation, differences in approach can appear arbitrary or politically motivated, undermining trust in institutions. By contrast, when authorities can point to shared principles, common practices and cooperative initiatives, they demonstrate that AI oversight is grounded in a broader normative framework.

Cooperation, therefore, plays a dual role: it improves the technical quality of supervision and reinforces the democratic legitimacy of regulatory institutions.

The Global Network for AI Supervisory Authorities (GNAIS), supported by UNESCO, can build on these lessons. It can serve as a focal point for cooperation, bringing together authorities with different mandates and capacities and providing them with tools to work together. By drawing on the experience of existing networks and adapting them to the specific challenges posed by AI, GNAIS can help establish a durable infrastructure for cooperation that supports supervisory authorities in their evolving roles.

Information-Sharing in Practice

Information-sharing is the foundation of cooperation. AI systems are complex, context-dependent and often opaque. No single authority will encounter every relevant failure mode or risk scenario. By sharing information, supervisory authorities can learn from each other's experiences, identify emerging patterns and respond to new challenges more quickly and effectively. Information-sharing can take many forms, each contributing to a more comprehensive understanding of how AI systems behave in different environments and to identify cross-border patterns.

One of the most valuable forms of information-sharing is the exchange of incident reports. When an AI system causes harm or behaves unexpectedly, the authority that receives the initial complaint or notification can document what occurred, how it was detected, and how it was addressed. Sharing this information with other authorities helps them prepare for similar incidents in their jurisdictions. For example, if a high-risk AI system used for credit scoring systematically disadvantages certain groups in one country, authorities elsewhere can be alerted to this possibility and review similar systems proactively. Over time, incident reports can reveal systemic issues in particular sectors, model architectures or deployment contexts, allowing authorities to address root causes rather than isolated symptoms. This is particularly important where AI systems are integrated into essential public services, where failures can have severe consequences for individuals and communities.[4]

Authorities can also share technical evaluation methods. Reviewing an AI system often involves analysing training data, model documentation, testing procedures and post-deployment monitoring arrangements. Authorities with more experience in certain techniques, such as robustness testing, bias assessment or explainability analysis, can share their methodologies, tools and lessons learned. This can happen through technical workshops, written guidance, or repositories of sample evaluation protocols. In data protection, authorities have long

[4] Marc Rotenberg and Christabel Randolph, 'The AI Red Line Challenge' (Tech Policy Press, 3 September 2024) <https://techpolicy.press/the-ai-red-line-challenge> accessed 28 November 2025.

shared templates for data protection impact assessments, privacy by design guidelines and enforcement summaries. Similar practices can support AI supervision by providing authorities with concrete tools and techniques they can adapt to their own legal and institutional contexts.

The Hiroshima AI Process launched the HAIP Reporting Framework to establish a standardised questionnaire for labs to describe, in a reasonably consistent format, how they manage safety and security for frontier models, so policymakers, peers, and the public can scrutinise and compare those practices. These reports typically include risk identification and evaluation; risk management & information security; model capabilities, limitations, misuse risks, and governance, oversight, and incident processes.

Another important category of information includes regulatory actions and guidance. When an authority issues a decision, adopts a code of practice or publishes interpretative guidance, these documents can inform authorities elsewhere, even if the legal frameworks are different. Shared knowledge of how others have interpreted concepts such as transparency, fairness or human oversight can help authorities refine their own approaches and avoid duplicating efforts. The International Consumer Protection and Enforcement Network, for instance, compiles summaries of significant consumer enforcement actions, which help authorities understand emerging risks and provide more consistent guidance across borders.[5]

Information-sharing can also include market observations. Supervisory authorities often observe trends that may not yet be the subject of formal investigations or guidance. These might include the rapid diffusion of certain AI services, changes in business models, or new forms of data collection and processing. Sharing these observations can help authorities build a more comprehensive picture of the AI landscape and anticipate developments before they become widespread. This is particularly relevant for frontier AI models and novel deployment contexts, where formal regulation may lag behind technological change.

Finally, information-sharing can include good practices and lessons learned. Authorities may experiment with different approaches to supervision, such as thematic reviews, sandboxes, or collaborative audits, and then evaluate their effectiveness. Sharing these experiences allows other authorities to learn from both successes and setbacks. It also helps avoid a situation in which each authority must independently develop and test supervisory methods. Summaries of these lessons help others refine their approaches.

Effective information-sharing can be facilitated through several practical mechanisms. Secure online services, hosted or coordinated by UNESCO, can provide a repository for incident summaries, evaluation methods and guidance. Access can be restricted to authorised officials, ensuring confidentiality where necessary. Shared repos

[5] 'Protecting Consumers Worldwide | ICPEN' <https://www.icpen.org/protecting-consumers-worldwide> accessed 28 November 2025.

tories maintained by UNESCO can store documentation, templates and case studies, ensuring that supervisory authorities in all Member States, regardless of resources, can access the latest materials. Regular coordination calls allow authorities to discuss emerging issues, share observations and plan joint initiatives. These calls can be organised by region, sector or theme, depending on needs. Periodic newsletters, briefing notes and internal reports can supplement these interactions, helping authorities maintain engagement and ensure that cooperation remains active.

Thematic workshops can support deeper engagement with specific issues. For example, workshops focused on high-risk AI in employment, health care, or financial services can bring together authorities, academics, civil society organisations and, where appropriate, providers, offering a service for learning and exchange. Bilateral and regional exchanges add further depth. A supervisory authority facing a particular challenge can seek advice from a peer with relevant experience, whether in the same region or elsewhere. Regional networks, such as NADPA or APPA, can focus on local challenges and build regional capacity.

Information-sharing must respect confidentiality. Some materials, such as details of ongoing investigations or trade secrets, cannot be shared widely. Authorities must therefore develop protocols that distinguish between public, shared and confidential information, and ensure that appropriate safeguards are in place. Anonymised summaries and high-level descriptions can often convey essential lessons without compromising sensitive details. Similarly, authorities must pay attention to legal constraints on data sharing, including restrictions on personal data transfers. These considerations are not unique to AI supervision; they have long been addressed in other regulatory domains and provide valuable precedents for AI supervisory cooperation.

Language accessibility is also critical. UNESCO's working languages can provide a starting point for documentation and exchanges, but AI supervision is global, and authorities work in many different languages. Providing translations of key documents and offering interpretation for meetings can help ensure that authorities with different language backgrounds can participate fully. Authorities can also contribute summaries and translations of their own materials, enriching the collective repository of resources.

By building on these mechanisms, the Global Network for AI Supervisory Authorities can create a robust infrastructure for information-sharing that supports supervision in a dynamic and rapidly evolving field. Shared information does not replace national decision-making, but it enhances it by providing a broader evidentiary base, reducing duplication, and supporting authorities with limited resources.

Beyond information-sharing, several additional mechanisms play a crucial role in supporting cooperation among supervisory authorities. These mechanisms help ensure that cooperation is structured, sustained and oriented toward concrete outcomes. Many of these mechanisms mirror the "enablers" for trustworthy public-sector AI identified by the OECD (governance structures, skills, tools and partnerships) and provide a practical way for supervisory authorities to operationalise those recommendations in their daily work.[6]

One foundational mechanism is mapping supervisory responsibilities. Although AI supervision is emerging as a common function, its allocation among institutions varies widely across countries. Some jurisdictions have a single, central authority; others distribute responsibilities among multiple sector-specific regulators; still others rely on coordination bodies that bring together existing agencies. A global map of supervisory responsibilities can help authorities identify their counterparts quickly and facilitate communication. Such a map would specify which authority or authorities are responsible for supervising AI in each jurisdiction, and in which sectors. It would also clarify how AI supervision interacts with other regulatory domains, such as data protection, competition, consumer protection and sector-specific regulation. UNESCO, working with regional networks and national authorities, is well-positioned to maintain such a map as part of its capacity-building work.

Annual surveys provide another essential mechanism. Surveys can gather information on supervisory activities, challenges and priorities, offering a snapshot of the state of AI supervision worldwide. They can capture quantitative data, such as the number of investigations, guidance documents issued or staff dedicated to AI, alongside qualitative insights about emerging issues and capacity needs. By consolidating survey responses, UNESCO can identify trends, highlight common challenges and showcase innovative approaches. Surveys also allow authorities to benchmark themselves against peers, helping them identify gaps in technical capability, legal authority or resources. The OECD's experience with questionnaires for its Coordinated Plan implementation reports shows that carefully designed surveys can provide rich data that inform both national policy and international cooperation.

Developing good practices is another important mechanism. Good practices are non-binding documents that distil experience and suggest practical approaches to supervision. They can cover topics such as documentation for high-risk AI systems, testing and validation methods, human oversight arrangements, complaint-handling procedures and post-deployment monitoring. Good practices can be developed through working groups composed of supervisory authorities and external experts. They can then be adapted by authorities to

fit their legal frameworks and institutional cultures. Over time, a collection of good practices can provide a shared reference framework that reduces uncertainty for both authorities and providers.

Engagement with external experts is also essential. Supervisory authorities cannot be expected to possess all the expertise required to evaluate complex AI systems in-house, particularly when resources are limited. Strategic engagement with academic researchers, technical experts, civil society organisations and professional bodies can help authorities stay abreast of technological developments and emerging risks. Engagement can take many forms: expert advisory panels, consultations on draft guidance, commissioned studies, joint workshops or collaborative research projects. The key is to structure engagement so that it enhances institutional capacity without compromising independence. Authorities must be mindful of potential conflicts of interest and ensure that external input complements, rather than replaces, their own judgment.

Annual meetings bring these mechanisms together into a coherent process. Regular gatherings, whether in person or online, allow supervisory authorities to review survey results, discuss emerging issues, share good practices and plan joint projects. Meetings can include closed sessions for confidential exchanges among authorities, as well as public sessions for engagement with civil society, industry and the broader public. Regular meetings help maintain momentum and build trust. They also provide opportunities to align activities with broader international processes, such as the implementation of the UNESCO Recommendation, the work of the OECD and the G7 Hiroshima AI Process.

A public-facing website also enhances transparency and accountability. Such a site can host information about the Global Network for AI Supervisory Authorities, including its objectives, members, activities and outputs. It can provide access to non-confidential good practices, survey summaries and public statements. It can also serve as a portal through which individuals, civil society organisations and providers can learn about supervisory expectations and contribute feedback. Transparency does not mean that all information must be public, but it does require clear communication about the role of supervisory authorities and how they cooperate. A well-designed website can help demystify AI supervision and enable the public to understand how authorities oversee AI systems.

Finally, cooperation must be sustained through a multi-year framework. One-off projects and ad hoc initiatives can be valuable, but they are unlikely to provide the continuity necessary for enduring cooperation. A multi-year programme, supported by UNESCO and aligned with the work of other international organisations, can provide a stable foundation for cooperation. Such a programme can set strategic priorities, allocate resources and establish timelines for key outputs, such as surveys, good practices and training sessions. It can also sup-

port evaluation and learning, allowing authorities to reflect on what has been achieved and where further work is needed. Sustained efforts build on previous work rather than restarting annually, ensuring that cooperation deepens over time.

These mechanisms (mapping, annual surveys, good practices, expert engagement, annual meetings, transparency initiatives and multi-year planning) are mutually reinforcing. Together, they help transform cooperation from a series of isolated interactions into a structured, strategic and sustained practice. In doing so, they strengthen the overall mission of ensuring safe, fair and transparent AI systems.

Conclusion

The emergence of AI supervisory authorities marks a significant development in the governance of digital technologies. As AI systems become more pervasive and powerful, societies need institutions capable of ensuring that these systems respect fundamental rights, promote fairness and remain accountable to the public. This chapter has argued that cooperation among supervisory authorities is essential if these institutions are to meet that challenge. Cooperation allows authorities to share knowledge, pool resources, coordinate responses to cross-border issues and develop consistent expectations for providers and deployers. It also enhances legitimacy by demonstrating that supervision is grounded in shared principles and informed by diverse experience.

The mechanisms described: information-sharing, mapping of responsibilities, annual surveys, good practice development, expert engagement, annual meetings, transparency and multi-year planning, offer practical ways to operationalise cooperation. They are not prescriptive; different jurisdictions will adapt them to their own needs and constraints. But they provide a framework within which authorities can work together, learn from one another and respond more effectively to the risks and opportunities presented by AI.

UNESCO's Supervising AI by Competent Authorities initiative provides an enabling context for these efforts. By offering a platform for cooperation, guidance on good practices and support for capacity-building, UNESCO can help ensure that supervisory authorities, especially those in the Global South and in smaller jurisdictions, are not left behind. OECD's analyses of AI governance in core government functions and of the implementation of the EU Coordinated Plan on AI provide additional analytical foundations, highlighting the importance of strong institutions, clear mandates and adequate resources. The G7 Hiroshima AI Process, with its focus on principles and codes of conduct for advanced AI systems, further underscores the need for robust supervisory capacity and for coordination among authorities across borders.

In the years ahead, the effectiveness of AI supervision will depend not only on the design of laws and institutions but also on the quality

of cooperation among supervisory authorities. If cooperation is robust, inclusive and grounded in human-rights principles, supervisory authorities will be better equipped to guide the development and use of AI in ways that support democratic values and social justice. If cooperation is weak or fragmented, AI governance risks becoming reactive, inconsistent and vulnerable to capture by powerful interests. The choice is not preordained. By investing now in cooperative mechanisms and by aligning supervisory practice with international frameworks such as the UNESCO Recommendation, the OECD reports, and the Hiroshima AI Process, the global community can help ensure that AI serves the public good and remains subject to the rule of law.

References

Council of Europe. 2024. Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law. Strasbourg, Council of Europe.

European Union. 2024. Artificial Intelligence Act. Brussels, Official Journal of the European Union.

Global Privacy Assembly. 2025. Rules and Procedures. London, Global Privacy Assembly Secretariat.

G7. 2023. Hiroshima AI Process Comprehensive Policy Framework, including Guiding Principles and Code of Conduct for Organizations Developing Advanced AI Systems. Hiroshima, Group of Seven.

OECD. 2025a. Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions. Paris, OECD Publishing.

OECD. 2025b. Progress in Implementing the European Union Coordinated Plan on Artificial Intelligence (Volume 1: Member States' Actions). Paris, OECD Publishing.

International Consumer Protection and Enforcement Network (ICPEN). 2024. Annual Report 2023–2024. The Hague, ICPEN Secretariat.
Randolph, C. and Rotenberg, M. 2024. The AI red line challenge. Tech Policy Press.

Rotenberg, M. 2006. The sui generis privacy agency: How the United States institutionalized privacy oversight after 9-11. SSRN.
Rotenberg, M. 2023. AI oversight. Foreign Affairs, November–December 2023. New York, Council on Foreign Relations.

Rotenberg, M. 2024. Human rights alignment: The challenge ahead for AI lawmakers. In: Werthner, H. et al. (eds), Introduction to Digital Humanism. Cham, Springer, pp. 453–462.
Rotenberg, M. 2025. AI policy (and democracy): A global perspective. Journal of Swiss Law (ZSR). Zurich, Schulthess.

Rotenberg, M. and Randolph, C. 2024. International frameworks for AI governance. In: Valcke, P., Forgó, N. and Pehlivan, C. (eds), AI Governance and Liability in Europe: A Primer. London, Edward Elgar Publishing.

UNESCO. 2024. Recommendation on the Ethics of Artificial Intelligence. Paris, UNESCO Publishing.

UNESCO. 2025. Supervising AI by Competent Authorities: Global Capacity-Building Initiative. Paris, UNESCO Publishing.

CONCLUDING REMARKS

This report forms part of the broader Supervising AI by Competent Authorities project and reflects the sustained engagement of supervisory authorities who are preparing to implement emerging AI governance frameworks in practice. The contributions presented here offer different perspectives on how institutions can organise their work, develop new forms of expertise, and cooperate effectively in a rapidly evolving technological landscape.

Throughout the project, authorities have emphasised the need for practical guidance that goes beyond legal texts. They have expressed interest in approaches that help them understand how supervision can be organised, how new responsibilities interact with existing mandates, and how evidence, expertise, and cooperation can be used to support effective oversight. This report aims to contribute to that effort by gathering analyses that support reflection and provide concrete considerations for supervisory practice.

Several broader observations emerge from this work. Supervisory authorities are gradually developing new capabilities to respond to the specific characteristics of AI systems, including their adaptability, cross-sectoral nature, and reliance on data-intensive processes. As they do so, they are beginning to make use of a wider set of methods: interpretative approaches that help them make sense of system behaviour, testing environments that allow for practical learning, and new forms of engagement with those who develop and deploy AI systems.

The report also highlights the continued importance of cooperation. Given the breadth of policy areas affected by AI, coordination between authorities within a country is essential for consistent and reliable implementation. Likewise, cooperation across countries supports shared understanding, avoids duplication of effort, and allows authorities to learn from each other's approaches. Initiatives such as the European Working Group of Competent Authorities on AI, NOBAREG, and UNESCO's Global Network of AI Supervisory Authorities play an important role in enabling this exchange.

The approaches discussed here are certainly not exhaustive, but they provide useful reference points for institutions developing or refining their own models of supervision. The intention is not to prescribe uniform solutions, but to offer material that can support authorities in identifying what is most suitable within their respective institutional and legal contexts..

Supervisory authorities will continue to play a central role in ensuring that the deployment of AI technologies aligns with legal requirements and broader public-interest objectives. As their responsibilities expand, the need for clear procedures, appropriate resources, and structured cooperation will become even more important. This report is intended as a contribution to that ongoing process, supporting authorities as they work to build supervisory practices that are coherent, informed, and responsive to technological and societal developments.

United Nations
Educational, Scientific
and Cultural Organization

Dutch Authority for Digital
Infrastructure
Ministry of Economic Affairs and
Climate Policy

Funded by
the European Uni