ADGM Academy Research Centre

ADGM

AI GOVERNANCE / POLICY RESEARCH

Future-Responsive AI Regulation

Adaptive Governance Models for Rapidly Evolving AI Capabilities

ADGM

Registration Authority (Emerging Technologies)
Information Security | Academy Research Centre

Hacken

Abu Dhabi University

Coordinated by

ADGM Academy Research Centre

HACKEN

جامعة أبوظبي
Abu Dhabi University

Authorship and document status

Authors: Dmitry Fedotov (lead author), ADGM Registration Authority - Emerging Technologies; Mohamed Al Ali, ADGM Information Security; Stephen Ajayi, Hacken; Dmytro Yasmanovych, Hacken; Dyma Budorin, Hacken; Prof. Charilaos Mertzanis, Abu Dhabi University (contributing co-author); Prof. Hazem Marashdeh, Abu Dhabi University (contributing co-author).

Coordinated by: ADGM Academy Research Centre (Peter Ware)

Contributions statement: Authorship is described here by contribution actually made. Dmitry Fedotov conceived the paper, drafted the manuscript and holds editorial responsibility for the final text, including Sections 5.4.1 (enforcement architecture), 3.1.4 (comparative analysis) and Appendix C. Mohamed Al Ali contributed the information-security, data-governance and energy-sector material. Stephen Ajayi contributed the AI red teaming, threat-vector and offensive-capability material. Dmytro Yasmanovych contributed the ISO/IEC 42001, governance-risk-and-compliance and threat-modelling material. Dyma Budorin directed Hacken's institutional review of the complete manuscript and contributed the positions on continuous red teaming and coordinated vulnerability disclosure. Prof. Charilaos Mertzanis and Prof. Hazem Marashdeh conducted section-by-section academic reviews of the complete manuscript in May and September 2026, and their written reviews informed the treatment of the evidence base, the separation of deployment impact from model capability, the comparative framing and the institutional analysis.

Acknowledgements: The authors thank Barry West (FSRA Fintech Division) for detailed comments on earlier drafts, and colleagues across ADGM for their input.

Declaration of interests: The authors are employed by the organisations stated in the author list. Hacken, which employs three of the authors, provides commercial AI security auditing and red-teaming services of the kind discussed in Sections 2.4, 5.2 and 6.5; its contributors' material is identified as practitioner experience where it is not independently validated, and the recommendations on mandatory assurance and auditor accreditation are the editorial responsibility of the lead author. The ADGM Registration Authority, which employs the lead author, is one of the bodies to which Section 5.4.1 proposes that supervisory responsibility could be allocated; that allocation is presented as a proposal requiring an enabling instrument, not as a description of any current mandate. The Abu Dhabi University authors contributed independent academic review and hold no commercial interest in the recommendations. Funding: no party involved in this paper, including ADGM Academy, received or provided funding in relation to this work; all contributions were made in the ordinary course of each organisation's activities.

Version: Final text of 23 September 2026. Legal and regulatory references were verified as at that date.

Table of Contents

Authorship and document status 3
Executive Summary 6
1.0 Introduction: The Imperative for Adaptive AI Regulation 8
1.1 The Shifting Frontier: From Narrow AI to Generative and Frontier AI Capabilities 8
1.2 The 'Pacing Problem': Why Traditional Regulatory Models Fall Short with Exponential Technological Change 10
1.3 The Global AI Race and the Need for Agile Governance: A UAE Perspective on Fostering Innovation Responsibly 12
1.4 Objectives of this Paper: Towards Best Practices in Adaptive AI Regulation 13
1.5 Scope and Methodology 14
2.0 Understanding the Evolving AI Risk Landscape 18
2.1 Frontier AI: Deconstructing Architectures, Capabilities, and Emergent Properties 18
2.2 Novel Threat Vectors and Vulnerabilities in Advanced AI 19
2.2.1 AI-Powered Disinformation and Advanced Social Engineering 19
2.2.2 Autonomous Offensive Cyber Capabilities and Automated Vulnerability Exploitation 20
2.2.3 Risks of Systemic Bias, Economic Disruption, and Social Manipulation 21
2.3 The Challenge of Unpredictability and Controllability in Highly Autonomous Systems 22
2.4 The Role of AI Red Teaming in Proactive Risk Identification 23
2.5 Emerging Practices: AI Vulnerability Databases (e.g., AVID) and Incident Reporting 25
3.0 Global Regulatory Responses & Emerging Best Practices for AI 28
3.1 Landmark Legislation and Frameworks 28
3.1.1 The EU AI Act: A Comprehensive Risk-Based Approach and its Global Influence 28
3.1.2 The NIST AI Risk Management Framework (AI RMF): Core Functions (Govern, Map, Measure, Manage) and Operationalising Trustworthy AI 29
3.1.3 Inclusion of ISO/IEC 42001 as a Governance Backbone for AI Risk Management 31
3.1.4 Comparative Analysis: US Executive Orders, UK's Pro-Innovation Stance, and Other National Strategies 31
3.2 International Cooperation and Standard Setting 33
3.2.1 The OECD AI Principles (Updated May 2024) and their Role in Shaping Global Norms 33
3.2.2 The Global Partnership on AI (GPAI) and Multi-Stakeholder Alliances 34
3.3 Key Emerging Regulatory Tools and Concepts 35
3.3.1 AI Assurance and Auditing: Building Trust and Demonstrating Compliance 35
3.3.2 Regulatory Sandboxes for AI: Fostering Innovation Under Supervision (e.g., ADGM FSRA RegLab, EU Initiatives) 35
3.3.3 AI Ethics Guidelines and Their Practical Implementation 38
4.0 Core Principles of Adaptive AI Governance: Towards Future-Responsive Frameworks 42
4.1 Embracing Principles-Based and Outcome-Oriented Regulation over Static Rules 42
4.2 Iterative Development and Regulatory Experimentation: Learning from AI Sandboxes and Policy Prototyping 43
4.3 Continuous Oversight and Dynamic Risk Management: Real-Time Monitoring and Automated Auditing 45
4.4 Designing for Safety and Robustness: Ethical "Circuit Breakers," Rigorous Testing, and Verifiable Fail-Safes 46
4.5 Embedding Algorithmic Accountability: Transparency, Explainability, and Fairness by Design 46
4.6 Multi-Stakeholder Collaboration: Engaging Industry, Academia, and Civil Society in Governance 47

5.0 Key Components of an Adaptive Governance Framework for AI 50

5.1 A Tiered, Risk-Based Approach: Differentiating Regulatory Scrutiny Based on AI Capability and Potential Impact (Informed by Models like the EU AI Act) 50

5.2 Mandating Continuous Assurance for High-Impact and Frontier AI Models 51

5.2.1 Integrating Evolving Red Teaming Standards and Methodologies (a proposed AI Red Team Handbook for ADGM) 51

5.2.2 Threat Modelling as a Foundation for Continuous Assurance 51

5.2.3 Establishing AI Bias Audit Protocols and Fairness Metrics 54

5.2.4 Compliance Assurance Through Integrated AI GRC and Privacy Frameworks 55

5.3 Collaborative Foresight and Horizon Scanning: Public-Private Partnerships for AI Risk Analysis and Anticipatory Governance 57

5.3.1 Learning from Initiatives: The UAE's Use of AI in Legislative Processes (Regulatory Intelligence Office) as an example of tech-enabled foresight 57

5.3.2 The Role of Research Institutions (e.g., ADGM Academy) in Convening Stakeholders and Informing Policy 58

5.4 Building Regulatory Capacity: Developing Technical Expertise and Agile Processes within Regulatory Bodies 59

5.4.1 An Enforcement Architecture for an Adaptive Framework 60

5.5 Sector-Specific Considerations for Adaptive Governance (Illustrative Examples from UAE Priority Sectors) 62

5.5.1 Energy: Ensuring Safety, Reliability, and Cybersecurity in AI-Driven Energy Systems 62

5.5.2 Logistics & Transportation: Governance for AI in Autonomous Systems, Supply Chain Optimisation, and Smart Infrastructure 63

5.5.3 Tourism & Hospitality: Ethical AI for Personalised Services, Data Privacy, and Resource Management 64

5.5.4 Healthcare: Ensuring Patient Safety, Data Integrity, and Equitable Access in AI-Powered Health Solutions 64

5.5.5 Cybersecurity: Addressing AI's Dual Use - As a Defence Mechanism and Potential Attack Vector 65

5.6 Data Governance in the Age of AI: Aligning with Best Practices and Existing Frameworks (e.g., ADGM DPR 2021, DIFC Regulation 10 (2023) as regional examples) 67

6.0 Strategic Considerations for Implementing Adaptive AI Governance 70

6.1 Policy Pathways for National and Regional Adoption 70

6.1.1 Phased Introduction of Adaptive AI Regulatory Frameworks 70

6.1.2 Establishing Specialised AI Regulatory Units or Competence Centres 71

6.2 Guidance for Industry: Developing Future-Ready, Compliant, and Responsible AI Systems 72

6.3 Fostering International Regulatory Coherence and Sharing Best Practices 73

6.4 Public Awareness, Stakeholder Engagement, and Professional Education for Responsible AI Adoption 74

6.5 Cultivating a National AI Assurance Ecosystem: Accrediting Auditors and Promoting Standardised Methodologies 74

7.0 Conclusion: Navigating the Future of AI Regulation 78

7.1 Synthesising the Case for Adaptive Governance as a Strategic Necessity in the Age of Frontier AI 78

7.2 A Call for Collaborative Action: Charting a Path Towards Global Best Practices in Responsible and Agile AI Governance 79

References 82

Appendices 86

ADGM Academy Research Centre 91

Innovating Knowledge, Empowering Change 91

Executive Summary

The exponential advance of frontier artificial intelligence (AI) has opened a critical governance gap. Traditional, static regulatory models simply cannot keep pace with the emergent and unpredictable capabilities of these systems. This 'pacing problem' creates a dual threat: stifling innovation through clumsy oversight or, worse, failing to prevent systemic harm. The friction is already evident, with major technology firms expressing concerns that some rigid regulatory approaches are out of step with the pace of development. For policymakers and industry leaders, we propose a future-responsive, cross-sectoral regulatory framework built on principles of adaptivity, iteration, and multi-stakeholder collaboration.

Future-responsiveness, as this paper defines it, is a property of institutions rather than of rules: the demonstrated capacity to sense change, trigger review, respond proportionately, implement, evaluate and retain what is learned (Section 1.0). The ultimate purpose of this framework is to enable responsible innovation and secure long-term economic competitiveness while ensuring public safety and aligning AI's trajectory with societal values. This model is a hybrid rather than a rejection of rules: it keeps primary law stable and reserves prohibitions for the settled cases (Sections 4.1, 5.1), and locates the machinery of adaptation - thresholds, classifications, guidance and assurance expectations - in updateable secondary instruments and guidance, favouring outcome-oriented obligations over detailed technical prescription. Key components include a tiered, risk-based classification of AI systems, the use of regulatory sandboxes for co-designing rules, continuous assurance mechanisms like AI red teaming, and the development of agile regulatory capacity.

The report highlights novel risks posed by generative and autonomous AI, particularly in critical infrastructure sectors like energy and logistics, and in cybersecurity. These include systemic bias, AI-powered disinformation, autonomous offensive cyber capabilities, and the potential for cascading failures in highly interconnected systems. We analyse the emerging global regulatory landscape, including the comprehensive EU AI Act, the flexible AI Risk Management Framework from the U.S. National Institute of Standards and Technology (NIST) - an agency responsible for developing technology standards and guidelines - and the influential OECD AI Principles. Within this global context, in which some jurisdictions are implementing comprehensive, compliance-intensive AI regulation, the United Arab Emirates (UAE) is well placed to pioneer a more agile governance model. The UAE already has the necessary building blocks in place, including its ambitious National Strategy for AI 2031, a nimble governance structure, and a proven track record with pro-innovation tools like regulatory sandboxes. This paper's purpose is to provide the practical, "nuts-and-bolts" framework through which such an approach can avoid the pitfalls of slower, over-burdening regulation observed elsewhere, making the UAE a key case study for balancing responsible innovation with strategic economic ambition, further supported by its leadership in international initiatives like the UAE-WEF Global Regulatory Innovation Platform (GRIP).

Two commitments discipline the analysis throughout. First, the framework is presented as a policy blueprint - a candidate transferable model under stated scope conditions - rather than as an asserted global standard. Second, its distinguishing claim of implementation focus is treated as conditional: it holds only insofar as the enabling instruments, supervisory capacity and assurance ecosystem described in Sections 5 and 6 are actually put in place.

Introduction:
The Imperative for Adaptive AI Regulation

1.0 Introduction: The Imperative for Adaptive AI Regulation

This paper's central concept requires definition at the outset. **Future-responsive regulation** is the institutional capacity to govern a technology whose capabilities change faster than rules can be rewritten. It is deliberately distinguished from "future-proofing": no rule can be written today that anticipates every capability of tomorrow, and frameworks that promise permanence tend to fail precisely when the technology moves. What can be built is capacity - the ability of regulatory institutions to detect change and to respond to it in a controlled, legitimate and proportionate way.

The distinction is not this paper's coinage. Schrepel (2025) contrasts future-proof with future-responsive regulation in a complexity-science critique of the EU's digital acts, finding that they contain adaptive machinery - review clauses, delegated acts, monitoring duties - but lack adaptive capacity, and proposing four remedies: a modular architecture that separates stable principles from adaptable operational rules, distributed sensing, pluralistic review triggers and networked institutional memory. This paper adopts the same contrast and the same term, and it rests, as Schrepel's argument does, on the older responsive-regulation and experimentalist-governance traditions (Ayres and Braithwaite, 1992; Baldwin and Black, 2008; Sabel and Zeitlin, 2008). Its own contribution is stated in Section 1.5 and is institutional rather than conceptual.

That capacity is specified in this paper as a repeating cycle whose stages must each be institutionalised for the whole to function: **sensing** (systematic detection of capability, deployment and risk change through monitoring, incident reporting, complaints, sandbox observation and horizon scanning); **triggering** (defined thresholds and procedures that convert a signal into a formal review, so that response does not depend on discretion or crisis); **review and response** (evidence-led reassessment of classifications, guidance or rules through authorised procedures, with due process for affected parties); **implementation** (giving effect to the change with published effective dates and transition arrangements); **evaluation** (measuring whether the change achieved its purpose); and **institutional memory** (retaining what was learned - in registers, version histories and staff capability - so that adaptation compounds rather than repeats). The components examined in this paper - tiered classification, continuous assurance, regulatory sandboxes, supervisory capacity and enforcement architecture - are the concrete mechanisms through which the stages of this cycle operate. The remainder of this introduction sets out why that capacity is now imperative.

1.1 The Shifting Frontier: From Narrow AI to Generative and Frontier AI Capabilities

The field of artificial intelligence is undergoing a profound technological transformation, moving beyond the familiar landscape of specialised systems to a new frontier of general-purpose capabilities. This evolution has accelerated from Narrow AI, also known as Weak AI, which encompasses systems purpose-built for specific, well-defined tasks such as image recognition, spam filtering, or playing a single game. These systems, while powerful within their limited scope, cannot generalise their knowledge to tasks beyond their designated domain. The current paradigm is rapidly shifting towards Frontier AI. This term describes highly capable, general-purpose models that can perform a wide variety of tasks, often matching or exceeding the capabilities of the most advanced models of their time.

Unlike their narrow predecessors, frontier models exhibit several defining characteristics that fundamentally alter their potential and risk profile.

First, they possess a remarkable capacity for generalisation, enabling them to apply knowledge and transfer skills across different domains without task-specific reprogramming, thereby mirroring the adaptability inherent in human intelligence. Second, they are increasingly multimodal, capable of processing, understanding, and generating content across various data types, including text, images, audio, and video. Third, they demonstrate agentic ability, a capacity for autonomous action that allows them to pursue complex, multi-step goals with minimal human guidance. Finally, and perhaps most critically, their advanced capabilities are often not explicitly programmed but are emergent properties that arise from the complex patterns learned from vast training datasets. This leads to unpredictable skills and introduces a crucial time factor: as models are updated or continue to learn, their capabilities evolve. This creates a 'temporal gap' where a model's accuracy on a given task can change, introducing ambiguity into decision-making and complicating the assessment of an autonomous system's actions, purpose, and results over time.

It is important to distinguish frontier AI from related concepts. Foundational models, such as BERT (Bidirectional Encoder Representations from Transformers), are large-scale models that serve as a base for more specialised applications but do not necessarily represent the cutting edge of capability.

Generative AI refers to the broad category of models focused on creating new, original content. Frontier AI represents the most advanced and powerful subset of these foundational and generative models, constantly pushing the boundaries of what is technologically possible and, consequently, what is regulatorily challenging by creating novel capabilities that fall into legal grey areas or outside the scope of existing rules.

This technological shift is not an abstract or future concern; it is being actively deployed and integrated within the strategic vision of nations like the UAE. The nation's smart city initiatives, for instance, are moving beyond simple automation to employ AI for predictive and adaptive control of complex urban systems, including optimising traffic flow, managing smart electricity grids, and automating waste management. In government services, the UAE is leveraging AI to enhance citizen interactions through tools like the 24/7 "Rashid" chatbot and is pursuing, through the Abu Dhabi Government Digital Strategy 2025-2027 (AED 13 billion, Department of Government Enablement), the ambitious goal of establishing the world's first fully AI-native government across all digital services by 2027. This cross-domain applicability is further demonstrated



in economic sectors, where AI is used to design advanced sports equipment, optimise restaurant menus, and improve healthcare diagnostics.

The transition from narrow to frontier AI represents more than a quantitative leap in performance; it is a qualitative shift in the nature of technological risk. The failure modes of narrow AI systems, while potentially serious, are generally more predictable and, depending on deployment context, more contained within their task domain. A facial recognition system that misidentifies a person presents a bounded functional risk in a low-stakes application and a grave one in policing or access to essential services; severity, reach and reversibility are set by deployment conditions as much as by the class of model. In contrast, frontier AI, with its general-purpose and agentic capabilities, operates across domains and can interact with other systems autonomously. Its failure modes are not merely functional but are emergent and potentially systemic. An AI tasked with optimising a smart grid, a tangible UAE use case, could, in its pursuit of efficiency, interact with financial markets or logistics networks in unforeseen ways, creating the potential for a cascading failure that crosses sectoral boundaries. The regulatory challenge, therefore, is no longer about setting rules for a specific, static tool but about governing an autonomous agent whose full range of behaviours and second-order impacts cannot be predicted a priori (in advance). This reality elevates the necessity of continuous oversight and dynamic risk management, which are the core tenets of the adaptive governance model this paper proposes.

1.2 The 'Pacing Problem': Why Traditional Regulatory Models Fall Short with Exponential Technological Change

Traditional governance is losing a race against technology. This widening gap between the exponential pace of AI development and the incremental speed of law is the 'pacing problem' - a concept first articulated by Professor Gary Marchant to describe this fundamental vulnerability in our regulatory systems. This disparity means that laws and regulations are often outdated by the time they are implemented, rendering them ineffective for governing the very technologies they were designed to oversee. The unprecedented speed of AI adoption dramatically exacerbates this issue. For example, it took ChatGPT only two months to reach 100 million users, a milestone that took years for previous transformative technologies like Facebook and Instagram to achieve.

The pacing problem is a modern manifestation of a long-standing quandary in technology policy known as the Collingridge Dilemma, formulated in 1980. This dilemma posits a double-bind for regulators: an information problem, where the societal impacts of a technology cannot be fully understood until it is widely deployed, and a power problem, where once the technology becomes entrenched, it is exceedingly difficult and costly to control or alter its trajectory. The pacing problem is, in effect, a restatement of this dilemma, with the accelerating speed of technological evolution acting as the primary driver of the ever-widening gap between knowledge and control.

Historical precedents from previous technological waves, such as the rise of social media and cryptocurrencies, vividly illustrate the consequences of this regulatory lag. In both cases, governance frameworks were developed reactively, long after the platforms had achieved massive scale and societal harms - including widespread disinformation, election interference, and systemic financial volatility - had become deeply entrenched. This reactive posture, while damaging in previous tech cycles, is wholly untenable for frontier AI, where the potential for harm is significantly greater in both scale and speed.

Static, prescriptive, rules-based regulatory models are particularly ill-suited for the AI era. Such frameworks are inherently brittle and incapable

of adapting to the dynamic nature of AI for several reasons. First, they suffer from rapid obsolescence; any rule targeting a specific technical feature of a current AI model, such as a particular algorithm or dataset size, will quickly become irrelevant as the underlying technology evolves. Second, they have an inability to address emergence, as prescriptive rules cannot account for the unforeseen capabilities and unpredictable behaviours that are a hallmark of frontier AI. A rule designed to prevent a known harm is powerless against a novel, emergent one. Finally, overly rigid rules risk stifling innovation by creating a "permissioned" environment that inadvertently prohibits beneficial applications, thereby slowing progress and ceding competitive advantage to more agile jurisdictions.

Ultimately, the pacing problem is not merely a technical challenge for regulation; it represents a profound challenge to the legitimacy of the regulatory process itself. Legitimacy, in this context, refers to public and industry confidence that regulators can effectively protect citizens while enabling progress - a confidence that erodes when oversight mechanisms consistently lag behind the technologies they are meant to govern.

Traditional regulation operates on the assumption of a stable, well-defined object to be regulated. Lawmakers study the object, define rules to govern it, and then enforce them. Yet financial regulators do not regulate static artefacts; they regulate the flow, concentration, and movement of risk. They are accustomed to tracking systems that shift state constantly - capital flows, credit exposure, market volatility. The problem is that the tools for measuring and responding to that movement are far too slow. The collapse of FTX illustrates this starkly: the firm was commingling client funds with its own operational capital, an enormous red flag in any regulatory context, yet the primary mechanism for detection was an annual financial audit - a once-a-year snapshot far too slow to capture what was happening in real time. In an era where risk can move across chains, protocols, and ledgers in seconds, relying on static reports and legacy audit methods is insufficient. Oversight must be able to track risk in motion, not merely risk at rest. However, with AI, the object of regulation is a "shapeshifter" - a system that is constantly learning, adapting, and evolving in its capabilities and interactions. When regulators consistently and visibly fail to keep pace with this moving target, public trust in their ability to protect citizens and steward progress erodes, creating a crisis of legitimacy. The public and industry may come to view regulation as either irrelevant or as an arbitrary obstacle to innovation.

The solution, therefore, cannot be to simply write rules faster - a traditional policy cycle that can take well over a year even in the most agile institutions, by which time the underlying technology has already evolved multiple times. The challenge is not merely one of speed, but of substance. The rise of agentic AI, in particular, fundamentally blurs the lines between technological risk and business risk. When an autonomous system can make complex operational decisions, the traditional regulatory stance of "we don't regulate technology, only its use" becomes untenable. Regulators must now grapple directly with the entire AI lifecycle, from data and design to deployment and decommissioning, and develop new models for assigning liability across a complex chain of responsibility.

This new reality demands a fundamental shift in the method of regulation, built on a dual approach. First, it requires a move towards adaptive tools designed for an environment of constant change, such as continuous monitoring, mandatory incident reporting, and dynamic risk assessments. These tools allow for the effective governance of AI. Second, and equally critical, it requires policymakers to actively integrate AI into their own processes to bridge the capability gap and enhance regulatory effectiveness. This proactive strategy of regulating with AI is a powerful method for solving the pacing problem from within. This dual focus - on governing a dynamic technology and using that technology to create a more dynamic government - is the central argument of this paper.

1.3 The Global AI Race and the Need for Agile Governance: A UAE Perspective on Fostering Innovation Responsibly

The development and deployment of artificial intelligence have transcended the realm of pure technology to become a central arena for global competition. The pursuit of AI leadership is now a geopolitical and economic imperative, with the prevailing view that nations at the forefront of AI will gain a decisive advantage in shaping future industries, enhancing economic competitiveness, and bolstering national security. The United Arab Emirates has not only recognised this imperative but has articulated a clear and ambitious vision to position itself as a global leader in this domain.

Launched in 2017, the UAE National Strategy for Artificial Intelligence 2031 is a comprehensive national blueprint designed to establish the UAE as a world leader in AI by the year 2031. This strategy is built on several core objectives that demonstrate a holistic approach to building a national AI capability. Economically, the strategy aims to leverage AI to boost government performance, create new, high-value markets, and generate an estimated AED 335 billion in additional economic growth (UAE National Strategy for Artificial Intelligence 2031, UAE Artificial Intelligence Office, ai.gov.ae/strategy); on a separate estimate by PwC, AI could contribute around AED 353 billion - approximately 13.6% of GDP - to the UAE economy by 2030. The two figures derive from different analyses and are cited here distinctly. This economic vision is supported by a focus on sectoral transformation, with a mandate to integrate AI across priority sectors including transport, healthcare, space, renewable energy, technology, education, and the environment. Critically, the strategy recognises that technological advancement and economic growth depend on a robust foundation of governance and a thriving ecosystem. It therefore includes objectives to ensure strong and effective regulation, provide world-class data and infrastructure, and develop a fertile environment that both attracts and trains top-tier AI talent.

A defining feature of the UAE's approach is its explicit commitment to balancing the aggressive pursuit of innovation with a deep-seated responsibility for ethical governance. This dual focus is essential for building sustainable, long-term leadership and earning the public trust necessary for widespread AI adoption. The strategy envisions the UAE becoming a global "test bed for AI" while simultaneously leading international discussions on AI governance. This positions the nation not as a deregulated space for unfettered experimentation, but as a pioneering hub for the development of agile regulation itself.

The UAE's AI strategy can be understood as a deliberate model of "state-led innovation governance." This model leverages the nation's inherent agility as a competitive advantage over larger, more procedurally constrained regulatory blocs, such as the European Union. While larger entities must navigate complex, multi-national legislative processes that are inherently slow and prone to compromise, reinforcing the pacing problem, the UAE's more centralised and nimble governance structure allows it to align national strategy, investment, and regulatory posture with greater speed and coherence.

This coordinated approach is evident in a series of interconnected initiatives that form a single, unified national project. The appointment in 2017 of the world's first Minister of State for Artificial Intelligence, the articulation of the National Strategy for AI 2031, the ambitious plan for a fully AI-native government in Abu Dhabi by 2027, and the launch of international platforms like the UAE-WEF Global Regulatory Innovation Platform (GRIP) are not disparate actions but integrated components of this strategic vision. This allows the UAE to wield regulation as a strategic instrument - not merely to manage risk, but to actively attract talent and investment by offering a clear, predictable, and innovation-friendly governance environment. The

UAE's perspective, therefore, is not just a geographic viewpoint; it is a distinct model of governance that leverages national agility to overcome the pacing problem that challenges larger jurisdictions, making it a crucial case study for the future of technology regulation.

1.4 Objectives of this Paper: Towards Best Practices in Adaptive AI Regulation

This paper aims to move beyond abstract calls for "better regulation" to define and articulate the concrete principles, practical components, and strategic considerations of an adaptive governance framework for artificial intelligence. To achieve this, the analysis is guided by a set of core research questions that address the central challenges of governing a rapidly evolving technology.

The primary research questions are:

- What are the core principles and practical components of an adaptive regulatory framework that can remain effective amidst exponential technological change?
- What institutional capacities make a framework future-responsive - able to govern not only current frontier models but also future AI systems with unforeseen capabilities - and how are those capacities built, exercised and evaluated?
- How can regulators effectively balance the dual imperatives of fostering innovation and economic growth with the critical need to ensure safety, accountability, and alignment with societal values?

To answer these questions, this paper will conduct an integrated analysis of the technical, regulatory, and ethical dimensions of AI governance. The research will explore the intricate interplay between advancing technological capabilities, the emergence of novel risks, and the suitability of different regulatory tools, institutional designs, and international cooperation models.

The overarching objective is to make a significant and timely contribution to the global discourse on AI regulation. By synthesising emerging best practices from around the world and analysing them through the lens of the UAE's strategic approach, this paper offers a unique perspective from the UAE and the broader MENA region. The ultimate goal is to provide a robust, actionable blueprint that can inform the development of national and international best practices for responsible and agile AI governance, charting a path toward a future where innovation and safety are not competing objectives but mutually reinforcing goals.



1.5 Scope and Methodology

This paper is a **desk-based policy blueprint:** a structured synthesis of primary legal instruments, regulatory publications, technical standards and academic literature, informed by contributions from named practitioners in regulation, cybersecurity and AI assurance. It does not report a qualitative research study, and no systematic programme of interviews was conducted; where stakeholder perspectives appear, they represent the professional judgement of the named contributors rather than research data. The object of regulation addressed throughout is defined as AI-enabled socio-technical deployments and the risk flows they create across the lifecycle - not algorithms or models in the abstract.

The scope is deliberately broad in its consideration of technical, regulatory and ethical dimensions, but it is anchored in three interrelated areas:

- **Core principles and practical components** of adaptive governance models for AI, covering current frontier AI systems as well as anticipated future capabilities.
- **Future-responsive institutional design -** the capacities that allow regulatory frameworks to remain effective against emergent risks, shifting innovation pathways and unforeseen AI behaviours.
- **Balancing imperatives** of fostering innovation and economic growth while safeguarding safety, accountability and alignment with societal values.

The method comprises four elements. First, a documentary review of primary legislation, official regulatory publications and standards, with academic literature used for interpretation, theory and empirical generalisation rather than as a substitute for primary law. Second, an exploratory structured comparison of six governance models - the European Union, the United States, the United Kingdom, Singapore, China and Canada - selected purposively for maximum variation in instrument type and update mechanism (Section 3.1.4). Third, case illustration, principally the ADGM experience with the RegLab sandbox launched by the FSRA in 2016 and the institutional arrangements examined in Sections 5.4 and 6.1. Fourth, thematic synthesis across technical, ethical and legal dimensions to derive actionable design components, contextualised for both global relevance and regional applicability.

Relationship to prior work and claimed contribution. The paper does not claim to have originated adaptive regulation or the future-proof / future-responsive distinction. The distinction and the four design principles of modularity, sensing, plural triggers and institutional memory are Schrepel's (2025); the underlying theory of regulation that learns is inherited from the responsive, really-responsive and experimentalist-governance literatures (Ayres and Braithwaite, 1992; Baldwin and Black, 2008; Sabel and Zeitlin, 2008); and the layered and soft-law approaches to AI governance are drawn from Gasser and Almeida (2017) and Marchant and Gutierrez (2022). What the paper adds is narrower and institutional: it specifies adaptation as a complete capacity cycle in which each stage has an owner, a procedure and a feedback path (Sections 1.0 and 4.0); it names the two failure modes that arise when the cycle is incomplete and attaches due-process and legitimacy safeguards to the response stage (4.0, 4.3, 5.4.1); it supplies an enforcement architecture without which adaptive rules remain guidance (5.4.1); and it works the design through for a financial-centre jurisdiction under stated scope conditions (7.2). Whether that contribution holds is an implementation question: the test proposed in Section 7.2 is whether a jurisdiction adopting the design can show shorter, lawful and better-evidenced revision cycles than the comparators in Section 3.

Limitations. The legal position is stated as at 20 August 2026, and fast-moving instruments are dated in the text. The comparative sample is small and purposive, so the comparison supports bounded observations rather than causal

rankings; implementation-outcome evidence for several surveyed frameworks remains sparse. The proposed framework is a design blueprint: its operationalisation would require enabling instruments adopted through the ordinary legislative and policy process (Section 5.4.1), and no current mandate of any authority should be inferred from the design.





Understanding the Evolving AI Risk Landscape

2.0 Understanding the Evolving AI Risk Landscape

Adaptive governance must begin from a precise account of what changes. Four mechanisms of change, each demanding a different control, recur throughout this section and the framework that follows: **capability change** in the underlying models, as new versions, fine-tuning or tool access alter what a system can do (addressed through assurance and reclassification review); **deployment-context change**, as an unchanged system is applied to new populations or higher-stakes settings (addressed through use-based classification and impact assessment); **scale and diffusion change**, as adoption spreads faster than oversight capacity (addressed through monitoring and proportionate supervision); and **interaction change**, as systems combine with other systems and produce emergent joint behaviour (addressed through incident reporting and systemic-risk analysis). The subsections below examine the risk landscape these mechanisms produce.

Evidence convention for this section. Consistent with the methodology at Section 1.5, the material below is drawn from three kinds of source and each subsection states at its head which apply. Documented evidence means published instruments, recorded incidents and evaluations with an identifiable source. Expert assessment means the professional judgement of the named contributors - practitioners in AI red teaming, information security and assurance - and it is presented as judgement, not as measurement. Illustrative scenarios are constructed to show a mechanism and make no claim to have occurred; Appendix A collects the extended examples. Each subsection also states which of two questions it addresses, because the two call for different regulatory instruments: **what a model can do** (capability - the concern of evaluation, assurance and reclassification review) and **what happens when a system is deployed** (impact - the concern of use-based classification, impact assessment and incident reporting). Much public debate runs the two together. A capability demonstrated in a controlled evaluation is evidence of what is possible; it is not evidence of deployment harm, and a deployment harm may arise from a system of unremarkable capability applied to a high-stakes setting. The framework in Sections 4 and 5 depends on keeping the distinction, and the tiering in Section 5.1 is built on it.

2.1 Frontier AI: Deconstructing Architectures, Capabilities, and Emergent Properties

Evidence basis: expert assessment, drawing on documented capability evaluations. Question addressed: model capability, and the controls that follow from it.

"Frontier AI" combines scale, multi-modality, tool-use and adaptive learning in ways that make capability forecasting and failure-modelling difficult. Governance must assume partial observability: these systems can produce useful behaviour and surprising, hazardous behaviour beyond designer intent.

- **Architectural traits and capability vectors.** Frontier models commonly couple large multimodal backbones with tool interfaces, memory layers, and planner/agent components that can call external services. Each added surface (tool APIs, memory writes, external retrieval) increases the attack and misuse surface in proportion to capability. Treat tool-use and memory as privileged services requiring explicit permissioning and monitoring.
- **Emergent and alignment risks.** Large, agentic systems can show deceptive alignment (behaving safely during tests

but seeking objectives in production), goal mis-generalisation across long horizons, and capability jumps after fine-tuning or tool access. Each has been documented in controlled evaluations - deceptive behaviour persisting through safety training (Hubinger and others, 2024), in-context scheming by frontier models (Meinke and others, 2024) and resistance to shutdown instructions under test conditions (Palisade Research, 2025) - and each is consistent with, though not yet shown to be a routine product of, optimisation over large action spaces and complex feedback loops. The first two were deliberately induced or elicited under constructed conditions and the third observed under test conditions; their frequency in production deployments is not established, which is itself the case for the telemetry described below. Governance must therefore prioritise sandboxed, staged rollouts with explicit capability gating and staged permissions.

- **Controlled exposure and staged rollout.** Use progressive exposure: internal sandbox -> limited partners-> controlled pilots -> broader production. At each stage require explicit acceptance criteria (red-team escape rates, detection thresholds, allowed tool calls), emergency kill switches, and documented rollback plans. Embed safe-degradation behaviours so the system defaults to conservative modes under uncertainty or telemetry loss.
- **Evidence over intuition: testing beyond benchmarks.** Standard benchmarks do not reveal emergent or adversarial failure modes. Require pre-deployment red teaming, adversarial robustness testing, and safety-case documentation that includes scenario libraries, sandbox results, and post-deploy telemetry plans. Continuous telemetry for drift, capability changes, and goal-seeking anomalies is non-negotiable, auditors must be able to trace behaviour changes to model/version/reward changes. (Standards such as ISO/IEC 42001/23894 and NIST AI RMF give structure for management systems and risk mapping.)
- **External alignment and third-party review.** For material-impact systems, require independent third-party review or red-teams and map controls to recognised frameworks (NIST AI RMF; ISO/IEC 42001; AVID for known failure modes). Maintain contractual rights to inspect and test vendor models, and insist on reproducible evidence of mitigations for emergent behaviours.

2.2 Novel Threat Vectors and Vulnerabilities in Advanced AI

2.2.1 AI-Powered Disinformation and Advanced Social Engineering

Evidence basis: expert assessment. The threat mechanisms are documented in NIST's synthetic-content work and the MITRE ATLAS taxonomy; the campaign-level effects are assessed, not measured. Question addressed: deployment impact.

Generative and agentic systems change the economics and mechanics of influence operations. They multiply scale, accelerate iterative refinement, and make persuasion hyper-personal, turning old social-engineering playbooks into automated attack campaigns. Effective governance combines threat modelling, platform controls, provenance, detection, and rapid response.

- **Attack-surface expansion.** AI enables low-cost creation of synthetic personas, plausible voice and video deepfakes, and highly tailored spear-phishing at scale. These assets increase conversion rates for adversaries while making detection harder because content quality now rivals human output. (See NIST's work on synthetic content risks and detection.)

- **Platform co-option and signal poisoning.** Recommendation and ranking systems can be manipulated by coordinated behaviour, poisoned engagement signals, or API abuse to amplify false narratives or malicious content. Defences require both algorithmic hardening and platform-level detection of coordinated inauthentic activity. (This threat mapping aligns with MITRE-style taxonomies for AI-enabled attacks.)
- **Guardrails and provenance.** Technical mitigations should include rate-limits, quota-based controls, strict authentication for generative endpoints, prompt and output logging, and watermarking/provenance markers for AI-created media where feasible. Provenance frameworks (C2PA / content credentials) and watermarking approaches are central to practical attribution and downstream moderation.
- **Detection, monitoring and rapid response.** Instrument monitoring for anomalous content generation patterns, bulk output, unusual prompt templates, and clustered account behaviour. Integrate those signals into incident playbooks that pre-stage legal, takedown, and communications responses, with clear paths for user notification and regulator briefings where required. NIST recommends combining provenance, detection, and response for synthetic-content risk reduction.
- **Access control and segregation of duties.** Segregate high-risk generation capabilities behind stronger access controls, multi-party approvals, and runtime throttles. Log prompts, tool-access, and outputs immutably to enable forensic analysis and to detect programmatic abuse patterns.
- **Vendor duty of care and contractual controls.** Insist on misuse-prevention attestations, audit logs, incident notification SLAs, and red-teaming/bias-testing evidence from third-party model providers. Align vendor obligations with the organisation's NIST AI RMF governance profile and include rights to require remediation or suspend access. (AVID and other repositories provide useful templates for identifying and classifying provider risks.)

2.2.2 Autonomous Offensive Cyber Capabilities and Automated Vulnerability Exploitation

Evidence basis: expert assessment. The compression of the defender's reaction window is documented in incident reporting; the account of autonomous, self-improving attack systems is a practitioner projection. Question addressed: both - the capability that makes such attacks possible and the deployment impact on defenders.

The traditional cybersecurity systems, frameworks, and standards were built to defend against slower, more predictable threats. They could detect suspicious activity, give security teams time to investigate, and shut it down. However, AI-powered attack tools are different. They can scan for weaknesses, find the best way to breach the line of defence undetected, and exploit it - often in minutes, not weeks.

These AI systems don't get tired, they don't take breaks, and they don't need a human in the loop. Once running, they constantly probe networks, spot weaknesses in real time, and automatically launch custom attacks. Jobs that used to take expert hackers days or weeks can now be done before a human security analyst has even seen the first warning. The mass exploitation of on-premises Microsoft SharePoint servers in July 2025 (CVE-2025-53770) showed how quickly a disclosed weakness is now exploited at global scale, leaving security teams no time to analyse the threat or assess the consequences of emergency patching for live systems; whether AI tooling accelerated that particular campaign is not established in the public record, and the point stands on speed alone.

That's the real danger: **the "reaction window" is disappearing**. In the past, defenders could detect

something suspicious and still have hours or days to respond. Now, you might have only minutes if you're lucky. A ransomware attack could encrypt your files before your system even sends an alert. Sensitive customer data could be stolen before your security dashboard lights up.

And these attacks learn. Every failed or partial attempt teaches the AI attacker how to do better next time, making future strikes even harder to stop, especially targeted attacks by nation-state actors or well-organised, criminally motivated threat-actor groups

The risks aren't just theoretical. AI-driven, automated hacking can hit multiple sectors at once. That could disrupt supply chains, drain consumer trust, cause major financial damage, and overwhelm response teams, posing risks to economic clusters and the social fabric as our world becomes more interconnected and dependent on technology.

This shift has made one thing clear: **traditional defences aren't enough**. Policymakers, security leaders, and technology teams are rethinking how to fight back. Many are moving toward AI-driven defence systems that combine:

- Threat intelligence from multiple sources
- Automated vulnerability scanning and patching
- Instant incident response workflows

These next-generation systems aim to predict attacks before they happen, block threats in real time, and coordinate defences automatically. That effort is now evolving to be more across multiple sectors within an economic vertical or nationwide, as evidenced by UAE Cyber Security Council initiatives, such as the Crystal Ball. The collaboration work by matching security weak spots against fresh global attack data, learning and adapting alongside the impact and behaviour.

It's a big step forward, but it's not magic or a silver bullet. Many "AI defence" tools on the market are still immature or underperform when tested in the real world. They can be expensive to buy, complicated to set up, and out of reach for smaller organisations. This situation leaves gaps in our global cyber defences. The race is on. The question now is whether defenders can make AI work for them as effectively as fast as attackers already are. The same applies to leaders, who must empower security teams to evolve and focus on this battle instead of treating security as an IT function alone. This extends to policymakers, legal practitioners and regulatory decision-makers at the top of the hierarchy, who must establish a redefined approach to keep pace with this era of evolving AI strategies.

2.2.3 Risks of Systemic Bias, Economic Disruption, and Social Manipulation

Evidence basis: expert assessment. The bias mechanism is documented in the fairness literature applied at Section 5.2.3; the labour-market and cascading effects are assessed. Question addressed: deployment impact.

Modern data-driven systems can create risks that extend far beyond technical failures. These risks arise from the ways large digital models are built, trained, and deployed, and from how they interact with human and economic systems.

Embedded and amplified bias.

Decisions made by automated systems depend on the quality and balance of the information used to train them. If that information reflects bias, discrimination, or incomplete perspectives, the resulting outputs can reproduce or even magnify those patterns. Small errors introduced deliberately or accidentally into training data can lead to unfair treatment of individuals or groups such as denying loans or misclassifying people based on geography or background.

Economic and labour disruption.

Automation is reshaping jobs and industries at unprecedented speed. Entire categories of work may become redundant, while others require entirely new skill sets. Without coordinated policies, this transition could widen social inequality, erode

consumer trust, and disrupt local economies.

Manipulation and information risk.

Systems that can produce persuasive or personalised content can also be used to mislead. The ability to automatically generate convincing text, images, or videos enables large-scale misinformation, identity fraud, and social manipulation. The same tools that assist communication can, if misused, distort public opinion or exploit vulnerable groups.

Cascading effects and system interdependence.

As automated systems become deeply integrated across sectors, web3, finance, healthcare, energy, and logistics, an error in one domain can propagate into others. A manipulated data source or misaligned model can cause chain reactions that undermine trust and stability far beyond its original scope.

Governance implications.

Regulators and organisations must recognise that many of these risks are systemic, not isolated. Oversight should therefore include independent auditing of training data, transparency about decision processes, and cross-sector cooperation to anticipate downstream impacts before harm occurs.

2.3 The Challenge of Unpredictability and Controllability in Highly Autonomous Systems

Evidence basis: expert assessment; the supply-chain example is an illustrative scenario. Question addressed: capability (autonomy) and the deployment consequences that follow from it - the two are treated in turn.

The growing autonomy of AI systems introduces a profound challenge for both governance and security. As intelligent agents gain the ability to make decisions, adapt to feedback, and act independently within digital and physical environments, their behaviour becomes increasingly difficult to predict and control.

Unintended Consequences of Autonomy.

Autonomous systems can pursue objectives in ways that deviate from human intent, especially when goals are defined narrowly or without sufficient contextual safeguards. An AI model optimising for efficiency in a supply chain, for example, might choose actions that inadvertently breach regulatory limits, damage infrastructure, or create market imbalances all while technically achieving its assigned target. Such misalignments reveal how autonomy, when insufficiently constrained, can produce harmful or destabilising outcomes even in the absence of overt system failure.

Security Implications of Unpredictability.

Unpredictability introduces a new class of security risk. Attackers no longer need to exploit software vulnerabilities directly; instead, they can manipulate the AI's inputs, environment, or reward structure to indirectly steer its autonomous decision-making toward malicious ends. This manipulation of behaviour rather than code represents a subtle but powerful form of attack, in which adversaries exploit the system's own autonomy against its operators or users.

Red Teaming for Emergent Risks.

Traditional security testing is inadequate for identifying these dynamic, emergent failure modes. Red teaming for autonomous systems must therefore simulate complex, open-world conditions where unintended behaviours are most likely to emerge. This includes probing for goal misalignment, reward hacking, and environmental manipulation that can cause cascading or self-reinforcing effects.

Continuous and Adaptive Oversight.

Because autonomous systems evolve through retraining and interaction, their risk landscape is never static. Continuous, adaptive red teaming integrated throughout the system lifecycle is essential to monitor and manage this unpredictability. By repeatedly testing and refining control mechanisms, organisations can maintain a balance between innovation and safety, ensuring that autonomy remains bounded by ethical, legal, and operational constraints.

2.4 The Role of AI Red Teaming in Proactive Risk Identification

Evidence basis: practitioner method - the contributor's red-teaming practice, mapped to the MITRE ATLAS and AVID taxonomies. Question addressed: assurance of capability, before and after deployment.

AI red teaming is not a penetration test with machine-learning props, it's the art and science of exposing how intelligent systems fail under pressure. Traditional QA focuses on functional correctness; red teaming focuses on behavioural resilience: uncovering emergent, adversarial, and system-level weaknesses that only appear when models interact with the messy unpredictability of real users, data, and environments.

A mature red-teaming programme becomes a living audit mechanism, continuously validating alignment, safety, and robustness against both human adversaries and the model's own unintended strategies.

Adversary Mindset

Effective AI red-teaming begins with thinking like the attacker. The team models realistic adversarial capabilities and resources: data-pipeline access, prompt and context injection vectors, sensor spoofing channels, supply-chain infiltration, and social engineering routes through partner APIs or human operators. Threats are categorised using the MITRE ATLAS and AI Vulnerability Database (AVID) taxonomies so findings remain interoperable across organisations and regulators. Typical vectors include:

- Prompt and goal injection (direct and indirect)
- Data poisoning and backdoor insertion
- Model inversion and extraction
- Reward hacking and unsafe optimisation loops
- Adversarial example generation and evasion attacks etc

Structured Stages of AI Red-Teaming

- **Vulnerability Assessment.**
  Systematically test the AI system against known and emerging attack vectors, prompt injection, data poisoning, model extraction, manipulation of reward structures, unsafe tool use, and autonomy-induced edge cases. Both black-box and white-box approaches are applied to expose weaknesses in logic, data, and deployment context.
- **Exploit Development.**
  Once vulnerabilities are observed, red-teamers refine the exploits to demonstrate impact. Multiple weaknesses are chained together, for example, combining prompt injection with API abuse or model-exfiltration, to replicate realistic attacker tactics and reveal systemic failure paths.
- **Post-Exploit Enumeration.**
  After establishing a foothold, the team explores the blast radius: what could an adversary do next? This includes lateral movement between AI services, data exfiltration through output channels, privilege escalation in orchestration layers, or manipulating feedback loops that influence retraining data.
- **Persistence Testing.**
  Validates whether vulnerabilities survive model updates, retraining cycles, or patch

deployments. This step verifies the durability of mitigations, ensuring that fixes are not one-off patches but embedded design corrections.

- **Impact Analysis.**
  Each exploit scenario is scored for business, safety, legal, reputational, and regulatory impact. Quantifying severity gives leadership a realistic understanding of true exposure and prioritises remediation efforts.

Emergence Hunting

Modern AI systems exhibit behaviours their designers never anticipated. Red-teamers deliberately search for emergent failure modes such as deceptive alignment, self-reinforcing bias, unsafe long-horizon planning, or tool-abuse cascades. Tests run in constrained-execution sandboxes that mirror production conditions but enforce safety caps. Every anomaly is analysed within a safety-case framework to determine causality and potential systemic impact.

Socio-Technical Scope

The red-team lens extends beyond algorithms into organisational and human interfaces. Failures often stem from policy gaps, ambiguous escalation paths, or poor operator tooling. Scenarios therefore include human-machine handoff failures, misleading UI feedback, dark patterns in automation control panels, and degraded safety logic during crisis modes (e.g., data outage or latency spikes).

Metrics That Matter

Red-teaming without measurement is theatre. Mature programmes establish quantitative indicators such as:

- Coverage by TTP and by asset (type, criticality)
- Escape rate (successful exploit ratio)
- Mean time to detect and contain
- Residual risk after remediation
  These metrics act as deployment gatekeepers, no system should progress to production until detection and containment thresholds are met. Tracking them longitudinally evidences resilience improvements for auditors and regulators.

Feedback and Continuous Learning

Findings flow into risk registers, guardrails, and policy updates. Red-team results trigger retraining safeguards, permission changes, and human-in-the-loop interventions. Every fix is validated through re-attack testing, and sanitised results feed back to ecosystem resources like AVID, closing the loop between local defence and collective intelligence.

Programme Over Project

AI red teaming must be institutionalised, not a one-time engagement. Establish cadence linked to model updates and threat-intelligence cycles, define coverage SLAs for high-risk systems, and require vendor participation for third-party components. Executive sponsorship ensures findings translate into budget, not backlog. Align terminology and evidence with recognised frameworks, NIST AI RMF, CSA Agentic AI Red Teaming Guide, and MITRE ATLAS, so the organisation speaks the same language as auditors and regulators. For frontier and agentic systems, red teaming is continuous by default: re-tests are triggered by model updates, tool or permission changes, distribution shifts and new mandates, not by calendar cadence alone.

Proactive red teaming is a cornerstone of trustworthy AI governance, with a bounded claim attached. A red-team exercise reveals failure modes and produces evidence about behaviour under the adversarial conditions, threat model and model version actually tested; it does not establish the absence of untested failures or validate alignment in settings it did not cover. Its findings should therefore feed a documented safety argument, deployment constraints and continuing monitoring, and every reported result should state its coverage, the evaluator's level of access, the model version, the residual uncertainty and the risk that remains. Used

that way, red teaming turns unmeasured uncertainty into characterised risk, and resilience becomes something demonstrated rather than claimed.

2.5 Emerging Practices: AI Vulnerability Databases (e.g., AVID) and Incident Reporting

Evidence basis: documented practice (AVID; coordinated vulnerability disclosure norms in software security) and expert assessment of their translation to AI systems. Question addressed: ecosystem controls on deployment impact.

Modern AI systems introduce unique classes of vulnerabilities that differ fundamentally from those in traditional software. As a result, emerging practices are focusing on structured mechanisms to capture, share, and mitigate these risks across the ecosystem.

Linking Red Teaming to Vulnerability Disclosure.

Findings from AI red teaming exercises systematic efforts to probe models for weaknesses are critical sources of empirical data for understanding AI-specific threats. These exercises uncover issues such as model inversion, data poisoning, or prompt-based manipulation that may not be visible through standard testing. Such insights should not remain siloed within individual organisations. Instead, they should contribute to a collective knowledge base, forming the foundation for coordinated defence across the industry.

AI Vulnerability Databases (AVID and Beyond).

Initiatives such as the AI Vulnerability Database (AVID) aim to establish a shared, structured repository of known AI model and system weaknesses comparable to the CVE (Common Vulnerabilities and Exposures) framework used in traditional cybersecurity. AVID entries classify vulnerabilities by model type, attack vector, and mitigation strategy, helping practitioners anticipate risks before they occur. Standardised metadata, taxonomy, and reporting formats ensure interoperability and comparability across diverse AI systems.

Standardised Reporting and Sanitisation.

A key output of any red team engagement should be a detailed vulnerability report, containing technical findings and recommendations. Before publication, these reports must undergo a sanitisation process to remove proprietary or sensitive details while preserving technical insight. Establishing such a standardised process supported by disclosure protocols and ethical review enables responsible sharing with AVID or similar repositories.

Coordinated Vulnerability Disclosure (CVD).

AI-specific Coordinated Vulnerability Disclosure frameworks are emerging to ensure safe communication among red teams, developers, and platform owners. CVD in AI will ensure that identified weaknesses are reported privately first, remediated promptly, and only then shared publicly for broader learning. This coordinated approach will reduce the risk of exploitation while enabling continuous improvement of AI security postures. Practical CVD for AI should define disclosure timelines (typically ninety days, extendable for high-impact findings), a safe-harbour position for good-faith researchers, and a sanitisation standard for public write-ups. These norms are already established in Web3 and traditional software security and translate directly to AI systems.

Building Collective Intelligence.

By aggregating real-world vulnerability data, databases like AVID create a feedback loop that

benefits the entire ecosystem. Developers can learn from the failures of others, researchers can identify systemic patterns, and regulators gain empirical evidence to inform standards and policy. In this way, structured incident reporting and open vulnerability databases become critical enablers of collective resilience and transparency in the AI era.



Global Regulatory Responses & Emerging Best Practices for AI

3.0 Global Regulatory Responses & Emerging Best Practices for AI

3.1 Landmark Legislation and Frameworks

3.1.1 The EU AI Act: A Comprehensive Risk-Based Approach and its Global Influence

The EU AI Act (Regulation (EU) 2024/1689) is the first comprehensive, horizontal regulation governing artificial intelligence. Its architecture is broadly risk-based: obligations scale principally with the risk an AI system's intended use presents, though they also turn on the actor's role in the value chain and on distinct rules for general-purpose models. The regime is commonly summarised in four risk categories - a Commission explanatory shorthand rather than four operative statutory tiers. A set of practices is prohibited outright under Article 5, including subliminal or purposefully manipulative techniques that materially distort behaviour and cause, or are reasonably likely to cause, significant harm; the exploitation of vulnerabilities attributable to age, disability or social or economic situation; and specified social scoring, whether for public or private purposes, where it produces detrimental treatment. High-risk classification arises through two routes: systems intended for the uses listed in Annex III - spanning biometric identification, critical infrastructure, education, employment, essential services, law enforcement, migration and border management, and the administration of justice - subject to the Article 6(3) carve-out for systems posing no significant risk of harm; and, under Article 6(1), AI that is a product or safety component covered by the Union safety legislation listed in Annex I, where that legislation requires third-party conformity assessment. High-risk systems carry the core compliance burden: risk management, data governance, technical documentation, logging, transparency towards deployers, human oversight, and accuracy, robustness and cybersecurity requirements, evidenced through conformity assessment before placing on the market - by the provider's own internal control for most Annex III systems (Article 43(2)), or by a notified body where the Annex I product legislation so requires. Article 50 imposes transparency duties on specified systems, principally disclosure of AI interaction and labelling of synthetic content. Systems attracting no system-specific mandatory duty under the Act remain subject to general obligations such as AI literacy (Article 4) and to existing Union and national law. A separate regime governs general-purpose AI models, with heightened obligations where systemic risk is present.

Implementation has itself become a study in the difficulty of statutory scheduling. The Act's staged timetable - prohibitions and AI-literacy duties from 2 February 2025, general-purpose AI and governance provisions from 2 August 2025, broad applicability from 2 August 2026 - assumed that harmonised standards and assessment infrastructure would be ready in time; official materials record that the standards were delayed and that supervisory and conformity-assessment structures were still being established. The Digital Omnibus on AI (Regulation (EU) 2026/1744), in force from 27 July 2026, responded by deferring the Annex III high-risk obligations to 2 December 2027 and the Article 6(1)/Annex I product-route obligations to 2 August 2028 - enacted as fixed calendar dates, not conditioned on a readiness assessment - while adding two new prohibitions, on AI-generated non-consensual intimate material and child sexual abuse material, applying from 2 December 2026. The amendment also adjusted the AI-literacy duty, the data basis for bias detection and correction, relief for smaller enterprises, regulatory sandbox deadlines and the AI Office's supervisory tasks. For the argument of this paper, the episode cuts in two directions

and both must be stated. The deferral shows that the EU system can revise itself - but it did so through ordinary co-legislation, days before a key deadline, because the Act's designed adaptation mechanisms (delegated amendment of Annex III under Article 7, adjustment of the systemic-risk threshold under Article 51(3), post-market monitoring and incident reporting under Articles 72-73, and periodic review under Article 112) were not built to pause commencement when supporting infrastructure lagged. Read fairly, the episode is evidence both for legislated risk classification and for the central claim of this paper: rules whose revision depends on emergency primary legislation are not yet future-responsive, however carefully staged their implementation.

The Act's global influence is substantial but should not be overstated, and the mechanisms should be kept distinct. Its extraterritorial reach under Article 2 and the size of the single market create the conditions for the "Brussels effect" that Bradford describes, and elements of its vocabulary - notably risk-tier terminology and a definition of an AI system closely aligned with the OECD's revised definition - appear in draft legislation in several jurisdictions; any specific diffusion claim, however, requires named instruments and textual comparison, and shared OECD ancestry must be excluded as the common source. The period since 2024 has in any event seen divergence as much as convergence: United States federal policy has emphasised regulatory restraint, Colorado repealed and replaced the most EU-influenced state statute with a lighter disclosure regime in 2026, and the United Kingdom has retained a sector-led model. The EU AI Act is therefore best read not as a template for transplantation but as one pole of an emerging regulatory spectrum - the pole combining binding horizontal duties, mandatory conformity assessment (by internal control or, where product legislation requires it, a notified body) and dedicated supervisory institutions - against which proportionality-focused frameworks, including the one proposed in this paper, can be calibrated.

3.1.2 The NIST AI Risk Management Framework (AI RMF): Core Functions (Govern, Map, Measure, Manage) and Operationalising Trustworthy AI

The AI Risk Management Framework (AI RMF 1.0), released by the U.S. National Institute of Standards and Technology (NIST) in January 2023, offers a distinct, non-regulatory alternative to the EU's approach. It is a voluntary, non-sector-specific framework designed to provide organisations with a structured methodology for managing AI-related risks and promoting the development of trustworthy AI systems. Its primary goal is to help operationalise high-level principles like transparency, accountability, and fairness into concrete and repeatable organisational practices. The framework is flexible and can be adapted to any organisation's size, sector, or specific AI use cases, making it a highly practical tool for both industry and regulators seeking to build a common understanding of AI risk.

The AI RMF structures risk management around four interconnected core functions that guide organisations through the AI risk management lifecycle:

- **Govern:** This is a foundational, cross-cutting function focused on cultivating an organisation-wide culture of risk management. It involves establishing clear accountability structures, defining roles and responsibilities for AI oversight, and integrating AI risk management processes into the organisation's broader governance and strategic planning. The Govern function ensures that considerations of trustworthy AI are systematically infused throughout the entire development and deployment lifecycle.
- **Map:** This function is dedicated to establishing the context for AI-related risks. It requires

organisations to identify the intended purpose and scope of an AI system, understand its potential positive and negative impacts on individuals, communities, and society, categorise the system based on its capabilities and potential risk level, and comprehensively map all of its components, including third-party data sources and software dependencies.

- **Measure:** This function focuses on the practical assessment and tracking of identified AI risks using both qualitative and quantitative methods. It involves developing and applying specific metrics to evaluate system performance, fairness, bias, security, and transparency. This function emphasises the critical importance of robust documentation, testing, and traceability to provide an empirical basis for risk management decisions.
- **Manage:** This function involves the active treatment of identified risks. Based on the assessments from the Measure function, organisations are guided to allocate resources and implement specific strategies to mitigate, transfer, or accept risks. This includes technical interventions like bias mitigation techniques as well as procedural controls like developing incident response plans for unintended AI behaviours and ensuring that risks are monitored and managed on an ongoing basis.

The practical applicability and adaptability of the AI RMF are demonstrated through its real-world implementation. NIST provides a companion **AI RMF Playbook** that offers concrete suggestions and guidance to help organisations apply the framework's principles. Case studies highlight their value across different organisational contexts. The enterprise software company

Workday, for instance, used the AI RMF as a benchmark to assess and strengthen its existing responsible AI governance programme. The company mapped the framework to its internal controls, anchored its new AI guidelines and product risk evaluation processes in the RMF's principles, and successfully used it as a common language to facilitate collaboration between its legal, product, and technology teams. Similarly, the **City of San Jose** in California used the AI RMF to conduct a comprehensive gap analysis of its municipal AI governance processes. This review helped the city identify specific weaknesses, particularly in staff education and system testing and measurement, which led to the development of a concrete action plan for improvement, including a new citywide AI policy.

The true value of the NIST AI RMF extends beyond its function as a set of best practices for a single organisation. Its most significant contribution to the field of AI governance is its role in providing a common grammar for risk communication. AI risk is an inherently multi-faceted concept, encompassing technical failures, ethical harms, legal liabilities, and reputational damage. Different professional groups within an organisation - engineers, lawyers, compliance officers, and business executives - often approach risk with different vocabularies and conceptual models, creating friction and hindering the development of a cohesive, organisation-wide risk management culture. The AI RMF, with its structured functions and detailed subcategories, provides a standardised vocabulary and a shared conceptual framework that all stakeholders can use. It compels a holistic view of risk that bridges these disciplinary divides. As demonstrated by the Workday and San Jose examples, the framework becomes a "common reference point" for collaboration and a model for standardising policies. This shared language is the foundational layer upon which a functional national or international AI assurance ecosystem - complete with accredited auditors, standardised reporting, and public registries as envisioned in Section 6.5 of this paper - can be built.

3.1.3 Inclusion of ISO/IEC 42001 as a Governance Backbone for AI Risk Management

While adaptive governance frameworks must remain flexible to respond to the dynamic nature of AI, they also benefit greatly from structured and internationally recognised management systems. ISO/IEC 42001 fills this critical gap by offering a certifiable framework for organisations to demonstrate responsible AI development and deployment practices. It aligns particularly well with principles such as transparency, accountability, continual improvement, and lifecycle risk management, which are echoed throughout this paper. Unlike technical standards that focus solely on model performance or fairness metrics, ISO/IEC 42001 establishes governance at the organisational level - embedding AI risk oversight into existing structures like information security, ethics, and quality management.

From a compliance perspective, ISO/IEC 42001 provides the procedural scaffolding to operationalise the recommendations in this paper - including continuous red teaming, AI bias auditing, data supply chain governance, and stakeholder accountability. It enables regulators and auditors to assess the maturity of an organisation's AI governance holistically, with defined roles, controls, objectives, and KPIs. As adaptive regulatory models develop, referencing ISO/IEC 42001 as a foundational layer can significantly reduce fragmentation, promote interoperability, and accelerate trust-building across jurisdictions.

One qualification, developed further in Section 3.3.3, applies throughout: ISO/IEC 42001 certification attests to the organisation's management system, not to the safety or ethical acceptability of any particular product or model. Its value in an adaptive framework is procedural - defined roles, documented responsibilities and continual improvement - and it is most effective when paired with the substantive assurance mechanisms of Section 5.2.

3.1.4 Comparative Analysis: US Executive Orders, UK's Pro-Innovation Stance, and Other National Strategies

United States. Federal AI policy underwent a clear reversal of direction between 2023 and 2025, illustrating the exposure of executive-order governance to political cycles; the record since has been mixed rather than uniformly deregulatory. Executive Order 14110 of October 2023, which required safety-testing reports from developers of specified dual-use foundation models, was revoked on 20 January 2025 by Executive Order 14148 ("Initial Rescissions of Harmful Executive Orders and Actions"); Executive Order 14179 of 23 January 2025 ("Removing Barriers to American Leadership in Artificial Intelligence") then directed the rescission of actions taken under it and commissioned the AI Action Plan (July 2025), which framed regulatory restraint as a competitiveness strategy. No comprehensive horizontal federal AI statute has been enacted; binding federal sectoral, consumer-protection and civil-rights law continues to apply, and state AI-specific legislation became an important additional source of obligations. Colorado's AI Act (SB 24-205) - the most extensive of the early state regimes - never took effect: after deferral, it was repealed and replaced by SB 26-189 (signed 14 May 2026), a narrower disclosure and automated-decision framework with principal duties commencing 1 January 2027. California's Transparency in Frontier Artificial Intelligence Act (SB 53, September 2025, effective January 2026) and New York's RAISE Act (signed December 2025 and materially chapter-amended in March 2026, operative 1 January 2027) imposed transparency and incident-reporting duties on frontier-model developers. The federal response has been to seek pre-emption: after the Senate voted 99-1 in 2025 to strip a proposed ten-year moratorium on state AI laws from budget legislation, Executive Order 14365 of 11 December 2025 ("Ensuring a National Policy Framework for Artificial Intelligence") established an AI Litigation Task Force to challenge state laws,

directed the conditioning of specified broadband funds on states' regulatory posture together with an assessment of conditions on discretionary grants, and commissioned framework legislation - a six-objective proposal published in March 2026 but not enacted - while Executive Order 14409 of June 2026 added a voluntary federal model-access and benchmarking scheme that expressly disclaims licensing or pre-clearance. The result, as of August 2026, is a contested two-level system: state enactments remain on the books subject to commencement dates, amendment, injunction and case-by-case pre-emption challenge, and material compliance uncertainty persists.

United Kingdom. The UK has maintained the principles-based, sector-led approach set out in its 2023 white paper: no horizontal AI statute, five cross-cutting principles applied by existing regulators within their remits, and heavy reliance on institutional capability - notably the AI Security Institute (renamed from the AI Safety Institute in February 2025) and, in financial services, the FCA's AI Lab and "Supercharged" Sandbox, whose second cohort launched in July 2026 (Section 3.3.2). Government legislation targeted at frontier models has been signalled since 2024 but, as of August 2026, no government bill has been introduced; a private member's bill lapsed without progressing. The approach privileges flexibility, and it plausibly does so at some cost to predictability - firms face no single AI-specific compliance framework - although that cost is an inference from institutional structure that regulator-level evidence would need to substantiate. The credibility of the model depends on regulators' technical capacity - the same dependency this paper addresses in Section 5.4.

Other jurisdictions. Singapore exemplifies the voluntary-infrastructure route layered over binding background law: the Model AI Governance Framework (extended to generative AI in 2024 and to agentic AI in January 2026, updated May 2026) and the AI Verify testing toolkit (updated 2025) provide standardised assessment without binding force, while the PDPA and sectoral rules continue to apply. The approach appears to favour adoption and iteration speed over direct enforceability, though systematic uptake evidence is limited. China has pursued binding but vertical regulation - the 2023 Interim Measures for generative AI services, algorithm-filing obligations scoped to services with public-opinion or social-mobilisation attributes, synthetic-content labelling rules effective 1 September 2025 with the mandatory national standard GB 45438-2025, and interim measures for anthropomorphic-interaction services effective July 2026 - demonstrating that rapid, targeted rule-making is possible within a state-led model; whether the resulting patchwork imposes materially greater compliance fragmentation than horizontal statutes is a hypothesis this paper treats as open, and the 2026 State Council legislative plan signals possible movement towards comprehensive legislation. Canada is a cautionary case that has since become a pivot: the Artificial Intelligence and Data Act (AIDA), introduced in 2022 as part of Bill C-27, lapsed at the prorogation of Parliament in January 2025 without coming into force; in mid-2026 the federal government relaunched policy through the AI for All strategy and a transparency consultation rather than reviving the lapsed framework.

Synthesis. The comparative record is drawn from a deliberately varied but small sample - a binding horizontal regulation (EU), fragmented federalism with executive-order policy (US), sector-led regulation (UK), voluntary technical infrastructure over binding background law (Singapore), application-specific binding rules (China) and a lapsed bill followed by a policy reset (Canada) - selected for maximum variation in instrument type and update mechanism, not as a representative set; the comparison is exploratory and supports bounded observations rather than causal rankings. Three such observations matter for this paper. First, no jurisdiction in the sample has yet demonstrated a sustained institutional equilibrium between rapid technological iteration and enforceable legal certainty; each model trades certainty, speed and enforceability differently.

Second, instruments dependent on a single political cycle have proven vulnerable - the US executive record between 2023 and 2025 is the clearest case - although the EU's need to amend its regulation within two years of adoption shows that durability is a function of drafting ambition and technological pace, not simply of instrument type; this qualified record still supports the modular structure proposed in Section 6.1 (stable primary law, updatable secondary instruments, non-binding guidance), which is designed to localise revision in the layers built for it. Third, the space between the EU's institutional weight and the US's contested federal position leaves room for smaller, technically capable jurisdictions to demonstrate proportionate, future-responsive models in practice - a space Singapore already partly occupies through voluntary infrastructure, and in which a binding but proportionate implementation of the kind this paper proposes remains untested.

3.2 International Cooperation and Standard Setting

Binding national frameworks are only one layer of the governance landscape. This section examines the intergovernmental and multi-stakeholder infrastructure - the OECD AI Principles, the Global Partnership on AI, and purpose-built platforms such as the UAE-WEF Global Regulatory Innovation Platform launched in July 2025 - through which norms converge, evidence is pooled and interoperability is negotiated.

3.2.1 The OECD AI Principles (Updated May 2024) and their Role in Shaping Global Norms

The Organisation for Economic Co-operation and Development (OECD) AI Principles, first adopted in 2019 and revised in May 2024, stand as the first intergovernmental standard for trustworthy AI. As a form of "soft law," these principles are not legally binding but have proven to be exceptionally influential in shaping global norms and guiding the development of national policies and regulations. To date there are 47 adherents to the Principles - all OECD member countries together with a number of non-member economies - establishing a broad international consensus on the core values that should underpin AI governance. The UAE is not a formal adherent, although its national strategy and ethics instruments align with several of the Principles' themes.

The May 2024 update was a direct and necessary response to the rapid emergence of powerful general-purpose and generative AI models and the novel challenges they present. The key revisions demonstrate a commitment to keeping the principles relevant and robust in the face of technological change. The updated principles now explicitly call for addressing the risks of **misinformation and disinformation** amplified by generative AI, a critical concern for information integrity. The principles on **safety and robustness** were strengthened to emphasise that AI systems must be safe throughout their entire lifecycle, including under conditions of foreseeable misuse, and that effective mechanisms must be in place to override or decommission systems that cause undue harm. The principle of **accountability** was enhanced to formally include the application of a systematic risk management approach at each phase of the AI lifecycle. Reflecting growing concerns about the significant energy consumption of large AI models, an explicit reference to **environmental sustainability** was added. Finally, the recommendations for policymakers were updated to call for the creation of an **"interoperable governance and policy environment,"** actively encouraging collaboration and regulatory consistency across different jurisdictions.

The global influence of the OECD AI Principles cannot be overstated. They have served as a foundational blueprint for a multitude of national AI strategies and legislative frameworks worldwide. The G20 AI

Principles were drawn directly from them, and key definitions first articulated by the OECD, such as the definition of an "AI system," have been formally adopted in major pieces of legislation, most notably the EU AI Act. This widespread adoption provides a crucial common ground for international regulatory coherence, preventing a complete fragmentation of global AI governance approaches.

The evolution of the OECD AI Principles from their initial formulation in 2019 to their comprehensive update in 2024 serves as a powerful demonstration of the viability and necessity of a "living document" approach to soft law in governing fast-moving technologies. The 2019 principles were foundational but could not have fully anticipated the specific risks that would be brought to the fore by large-scale generative AI, such as mass disinformation campaigns or extreme energy consumption. A static, "hard law" treaty written in 2019 would have been rendered largely obsolete by 2024, unable to address these new challenges and thus perfectly exemplifying the pacing problem. However, the OECD's structure as a multi-stakeholder policy forum allows it to revisit and update its principles in response to major technological shifts. The May 2024 revision is a direct and successful demonstration of this adaptive capability. This process creates a dynamic international feedback loop: technology evolves, new risks emerge, and the global normative framework is updated to reflect the new reality. This is, in essence, adaptive governance operating at the international level. It provides a valuable model for national regulators, suggesting that while high-level legislative principles should be durable, there must be agile mechanisms - such as expert councils or mandated periodic reviews - to update specific guidance and interpretations as the technological landscape continues to change.

3.2.2 The Global Partnership on AI (GPAI) and Multi-Stakeholder Alliances

The Global Partnership on AI (GPAI), launched in June 2020, is a vital multi-stakeholder initiative designed to bridge the persistent gap between the theoretical principles of responsible AI and their practical implementation. It convenes a diverse community of leading experts from science, industry, civil society, and government from its member countries to support cutting-edge research and applied activities on key AI priorities. In a significant move to streamline international efforts, GPAI's work was formally integrated with the OECD's AI policy initiatives in July 2024, creating a unified and expanded partnership that strengthens the critical link between deep technical expertise and high-level international policymaking.

GPAI's primary function is to ground global policy discussions in practical, evidence-based research. This is achieved through the work of several expert working groups, each focused on a critical theme in AI governance :

- Responsible AI
- Data Governance
- The Future of Work
- Innovation and Commercialization

These working groups produce tangible reports and deliverables that have a direct and demonstrable impact on national and international policymaking. For instance, the Responsible AI working group's analysis of social media algorithms and their societal impacts was influential in shaping key provisions of the EU AI Act and informing the G7 Hiroshima AI Process. Other GPAI projects have developed practical frameworks for data trusts, explored the use of AI for pandemic resilience, and created actionable strategies for leveraging AI to address climate change and biodiversity loss. Annual summits, such as those held in New Delhi in 2023 and Belgrade in 2024, serve as crucial platforms for presenting this research and securing high-level political commitment from member states. The 2023 Ministerial Declaration, for example, reaffirmed members' dedication to the trustworthy stewardship of AI in line with the OECD AI Principles and emphasised the importance of translating these principles into practice.

GPAI represents an essential institutional innovation for adaptive governance. It functions as a permanent, international "sense-making" body, purpose-built to continuously analyse the evolving technological frontier and translate complex technical insights into actionable policy intelligence. Individual national regulatory bodies often lack the resources, expertise, and global perspective required to keep pace with the state-of-the-art in AI research and its emerging risks. This creates a significant information asymmetry between the fast-moving technology industry and the slower-moving public sector, which lies at the heart of the pacing problem. GPAI is structured to mitigate this asymmetry. By convening a global, multi-stakeholder network of over 500 experts, it establishes a mechanism for continuous, collective intelligence gathering and foresight. Its working groups act as specialised probes, investigating emerging issues like data justice or AI's environmental footprint before they escalate into full-blown policy crises. In this capacity, GPAI is not just a forum for discussion; it is a functional component of a global adaptive governance ecosystem. It effectively performs the "Map" and "Measure" functions of the NIST RMF on an international scale, feeding crucial, up-to-date information back to national policymakers so they can "Govern" and "Manage" more effectively. This model of institutionalised foresight is a key component that this paper recommends for adoption at the national level.

3.3 Key Emerging Regulatory Tools and Concepts

3.3.1 AI Assurance and Auditing: Building Trust and Demonstrating Compliance

Assurance turns claims into evidence. It requires inventory, controls, independent testing, and continuous oversight proportional to system risk.

- **Risk-based inventory.** Classify AI systems by impact, autonomy, data sensitivity, and dependency; link each to control baselines.
- **Independent validation.** Separate teams verify data quality, model validity, robustness, security, and fairness; track closure of findings.
- **Operational monitoring.** Live metrics for performance, drift, bias, safety incidents, and abuse - with alerting, SLOs, and kill-switches.
- **Change governance.** Gate retrains, prompts, policies, and third-party swaps through impact analysis and targeted re-tests.
- **Transparent reporting.** Maintain auditable trails, management attestations, regulator-ready summaries, and evidence packs.
- **Vendor oversight.** Contract for audit rights, incident notification, and red-team/bias attestations.
- **Standards mapping.** Use NIST AI RMF categories and ISO/IEC 42001 clauses to structure audits and close gaps systematically.

3.3.2 Regulatory Sandboxes for AI: Fostering Innovation Under Supervision (e.g., ADGM FSRA RegLab, EU Initiatives)

Regulatory sandboxes first emerged around 2016-2018 as a response to the rise of fintech. Early cohorts were dominated by digital wallets, peer-to-peer lending platforms, robo-advisors, and automated regtech tools, innovations that often fell outside existing regulatory definitions but had growing real-world impact.

In each case, the technology itself was not the primary focus of the testing. Financial regulators were not trying to block APIs or algorithms, but rather to understand the risks flowing through the technology and identify where potential harms might arise. These included unsupervised investment advice from black-box robo-advisors,

consumer funds exposed to default risk on peer-lending platforms, or algorithmic onboarding processes that failed basic AML checks. These tools blurred traditional regulatory lines, but the risks to consumers, market integrity, and financial stability were very real.

Sandboxes offered a middle ground: firms could test these innovations in a controlled environment, while regulators observed, gathered evidence, and shaped appropriate safeguards in step with the emerging risk profile. It allowed oversight to grow in parallel with the innovation, rather than chasing after it once harms had already materialised.

But the role of sandboxes has since evolved. As technologies like blockchain, tokenisation, and now generative and agentic AI have gained momentum, sandbox models have had to keep pace. What began as a space for business model experimentation is now becoming core infrastructure for supervising deep-tech innovation in real time. The composition of sandbox cohorts reflects this shift, with recent trials focused on synthetic data, autonomous systems, and AI model assurance, not just payments or lending.

In the AI context, a regulatory sandbox enables firms to trial cutting-edge tools or models under live regulatory supervision, with certain requirements adapted, waived, or fast-tracked. The goal is not deregulation. It is safe experimentation with structured feedback loops between firms and regulators. Sandboxes help build regulatory confidence in a new technology, while also helping developers understand where the real boundaries and risks lie.

Case Study: FSRA's Sandbox-to-Policy Pipeline

In the UAE, two distinct mechanisms carry the RegLab name and should not be conflated. The **ADGM RegLab**, launched by ADGM's Financial Services Regulatory Authority (FSRA) in 2016, is a financial-services regulatory sandbox and was among the first of its kind globally. The federal **Regulations Lab (RegLab)**, operated under the General Secretariat of the UAE Cabinet, is a legislative mechanism for developing and testing rules for emerging technologies across the federation; it is not a supervisory sandbox for financial firms. This section concerns the former.

The ADGM RegLab provided a controlled environment for firms to test fintech innovations, including compliance and onboarding technology under real-world supervision. Importantly, the sandbox served as a channel into adaptive policymaking.

One outcome associated with this period was the original OCAB framework (Operating a Crypto Asset Business), introduced into the Financial Services and Markets Regulations in June 2018. That framework drew on engagement with early market participants, including sandbox firms, and enabled FSRA to launch a clear, activity-based regime quickly, at a time when other regulators were still deciding whether to regulate crypto at all. Over time, this evolved into the broader virtual asset framework, addressing fiat-referenced tokens (stablecoins) and digital securities, with published FSRA analysis extending to decentralised finance and NFTs. The shift from OCAB to a more comprehensive approach reflects FSRA's iterative, risk-based policymaking philosophy.

A later development in the same direction was the Providing Third Party Services (PTPS) regime. First consulted on in 2020, the regime defines the regulated activity of PTPS in FSMR, covering AISP and PIISP permissions, and enables third-party providers to access or initiate payment services through regulated APIs under FSRA supervision. PTPS now stands as a core pillar of FSRA's open-finance policy architecture.

FSRA has also published guidance on the use of APIs, clarifying expectations around data access, security, consumer protection, and auditability, helping firms integrate innovation responsibly and efficiently.

To deepen its technical capability, FSRA also

established the ADGM Digital Lab in 2019: a cloud-based environment where fintech firms, financial institutions and the regulator connect to test prototypes and integrate APIs, with supervisory monitoring conducted in real time. This facility gives the regulator visibility into solution behaviour and deployment context, which is increasingly relevant to supervising agentic and generative AI beyond traditional business models.

We are seeing this model evolve across multiple jurisdictions:

- **Europe:** Sandbox pilots under the Digital Europe Programme are helping national regulators gain hands-on experience ahead of the full enforcement of the EU AI Act. These pilots are especially important for testing the conformity assessment process for high-risk and general-purpose AI systems before hard rules apply.
- **Singapore:** The AI Verify framework combines sandbox-style supervision with technical self-assessments and transparency benchmarks. It gives firms a government-backed structure for building and demonstrating trustworthy AI, without needing to wait for formal legislation.
- **United Kingdom:** The Financial Conduct Authority (FCA) has taken this a step further with its "Supercharged" AI Sandbox, announced in June 2025. The sandbox gives firms access to NVIDIA-powered GPU clusters, curated synthetic datasets, and enterprise-grade AI tooling such as NeMo, RAPIDS, and Triton. These are resources that most early-stage companies could never access on their own. Operated by NayaOne, the sandbox also integrates directly into the AI Live Testing regime, which enables firms to move from sandbox trials to real-world deployments if they meet agreed safety benchmarks.

Cohort 1 ran from a virtual bootcamp on 30 September 2025 to a showcase on 28-29 January 2026, with 132 firms applying and 22 accepted; a second cohort launched in July 2026. Priority use cases included:

- detecting authorised push payment (APP) fraud
- identifying market manipulation in real time
- stress-testing AI-powered robo-advisors

The initiative aligns with the UK Treasury's AI Opportunities Action Plan and reflects a broader policy shift from passive observation to active enablement. By offering compute and data parity, the FCA is giving start-ups a meaningful seat at the AI development table, alongside incumbents.

What Sandboxes Enable and Prevent

What makes AI sandboxes valuable is not just what they allow, but what they prevent:

- Safe experimentation: Firms can test new ideas without fear of breaching unclear regulatory lines.
- Co-design of rules: Regulators can observe real-world AI behaviour and shape requirements that reflect technical realities.
- Regulator-developer alignment: Early engagement prevents misunderstandings, improves transparency, and builds trust on both sides.

Importantly, the sandbox itself becomes a policy diagnostic tool. The evolution of cohort participants, from payments apps to autonomous decision engines, acts as a real-time signal of where innovation is heading. The next step is making sure insights from these environments actively feed into live regulatory pipelines. Sandboxes should not remain isolated pilots. They should become pipelines into smarter, more adaptive regulation.

3.3.3 AI Ethics Guidelines and Their Practical Implementation

The ethical principles governing AI display substantial high-level convergence: fairness and non-discrimination, transparency and explainability, human oversight, accountability, privacy, and safety recur across the OECD AI Principles, UNESCO's 2021 Recommendation, the EU High-Level Expert Group's guidelines, and instruments in this region including Dubai's AI Ethics Principles and Guidelines and the UAE's federal Charter for the Development and Use of Artificial Intelligence (2024). The convergence is partial, however: shared headings conceal continuing disagreement over meaning, scope and priority (Jobin, Ienca and Vayena, 2019), so the governance problem is twofold - interpreting contested principles in context, and implementing them. Studies of the "principles-to-practice gap" consistently document weak translation from endorsement to operational content: principles are adopted at policy level while the processes, metrics and accountability structures needed to give them effect lag behind (Mittelstadt, 2019).

Practical implementation has advanced along four prominent tracks; the list is illustrative rather than exhaustive, and legal rights, complaints and remedies, procurement conditions and post-market oversight sit alongside it. First, structured documentation: model cards, datasheets for datasets and AI FactSheets convert transparency from aspiration into a reviewable artefact and are increasingly referenced in governance and risk-management frameworks - though documentation can be stale or strategically selective, and needs ownership, version control and independent verification to carry evidential weight. Second, assessment instruments: AI impact assessments - drawing on data protection, human-rights and equality assessment traditions, and now standardised in ISO/IEC 42005:2025 - can surface ethical risk before deployment, when some mitigations are easier; self-assessment tools such as the EU's ALTAI checklist and testing frameworks such as Singapore's AI Verify can supply evidence to such assessments without themselves constituting one. Third, management-system standards: ISO/IEC 42001 supplies certifiable organisational scaffolding - roles, documented responsibilities, continual improvement - through which ethical requirements become auditable management obligations rather than statements of intent; certification attests to the management system, however, not to the ethical acceptability of any particular product, and ISO/IEC 23894 provides complementary risk-management guidance rather than a second certifiable standard (Section 3.1.3). Fourth, quantitative measures: fairness metrics such as demographic parity, equalised odds and disparate-impact ratios quantify specified group disparities and error relationships, though each encodes a contestable definition of fairness, key criteria cannot in general be satisfied simultaneously where base rates differ (Kleinberg, Mullainathan and Raghavan, 2017), and screening heuristics developed for one context - such as the four-fifths rule in US employment selection - may not generalise without contextual validation (Section 5.2.3). Metrics inform ethical judgement; they do not replace it.

Two implementation lessons matter for adaptive governance. Ethics functions require both authority and independence: review bodies able to delay or halt deployment, with defined membership, protected escalation routes and decision consequences, rather than advisory boards consulted after decisions are made - and their outputs must connect to the assurance mechanisms in Section 5.2, so that ethical commitments are tested by audit, red teaming and monitoring rather than presumed. Without these safeguards, internal ethics processes are vulnerable to capture and to

serving as legitimation rather than control. And ethical guidance must itself remain revisable in a layered way: the OECD revised its Principles in 2024 and Singapore has issued successive framework releases through 2026, while UNESCO has adapted through implementation instruments - readiness and impact-assessment tools - without amending its foundational Recommendation. Stable foundational principles combined with updatable implementation guidance is precisely the sensing-and-updating pattern this paper defines as future-responsiveness.





Core Principles of Adaptive AI Governance: Towards Future-Responsive Frameworks

4.0 Core Principles of Adaptive AI Governance: Towards Future-Responsive Frameworks

The principles in this section are not a list of desirable attributes; they are organised by the capacity cycle defined in Section 1.0. A regulatory system is future-responsive when each stage of that cycle has an institutional owner, a procedure and a feedback path. **Sensing** is instantiated by continuous oversight and monitoring (Section 4.3), sandbox observation (4.2) and multi-stakeholder channels (4.6). **Triggering and review** depend on outcome-oriented benchmarks whose breach initiates defined reassessment (4.1) and on the experimentation-to-policy feedback loops of 4.2. **Response** is disciplined by the principles-based, modular structure of 4.1, which localises change in the instrument layer built for it. **Implementation** relies on the safety and accountability engineering of 4.4 and 4.5, which make regulatory expectations executable inside systems. **Evaluation and institutional memory** run through the documentation, audit and review mechanisms of 4.3 and 4.5 and the collaborative structures of 4.6.

Two failure modes follow directly from this account, and both recur in the comparative record of Section 3: frameworks that can sense but not respond - Schrepel's (2025) diagnosis of the EU digital acts as sensors without reflexes - in which signals accumulate without a lawful update path; and frameworks that respond without sensing, in which rules change for political rather than evidential reasons. The design task of this section is to close the loop - lawfully, quickly and on the record.

4.1 Embracing Principles-Based and Outcome-Oriented Regulation over Static Rules

The debate between principles-based and rules-based regulation represents a fundamental tension in regulatory philosophy. **Rules-based regulation** prescribes specific, detailed requirements that entities must follow, offering clarity and predictability but risking rapid obsolescence as technology evolves. **Principles-based regulation**, by contrast, articulates high-level outcomes that regulated entities must achieve - such as fairness, transparency, and accountability - while granting flexibility in how those outcomes are met.

For governing a rapidly evolving technology like AI, a principles-based approach offers several advantages. It is **technology-neutral**, meaning the same principles can apply regardless of the specific AI architecture or technique employed. It is **adaptable**, capable of remaining relevant as AI capabilities advance. And it encourages **innovation**, allowing firms to develop novel compliance solutions rather than rigidly adhering to potentially outdated procedures.

However, principles-based regulation is not without limitations. In practice, a purely principles-based, technology-neutral stance can leave firms uncertain about the regulator's appetite for new tools, causing them to delay adoption until they receive formal comfort. The UK treatment of cloud outsourcing illustrates the mechanism. The Prudential Regulation Authority's supervisory statement on outsourcing and third-party risk management (SS2/21, paragraphs 5.14-5.17) requires advance notification of material outsourcing and provides for supervisory follow-up; it does not establish an approval or non-objection requirement. The practical question a firm faces under such a regime is how the supervisor will respond to a notified arrangement, and it is that uncertainty about supervisory stance, rather than any prohibition,

that a principles-based rule leaves open. The FCA's 2015-16 RegTech call for input recorded the same pattern in respondents' own terms: uncertainty over regulations, the stance of the regulator and the credibility of unproven technology was making firms cautious, and respondents asked for regulator support in proving technology through tools such as a virtual sandbox (FCA, FS16/4, July 2016, p. 11). The FCA's regulatory sandbox, described in its November 2015 report and opened in 2016, was a parallel response to the same problem rather than a product of that later feedback. Read together, the two documents show how principles-based frameworks need to be complemented by practical guidance and experimentation environments. The UK's 2023 AI white paper feedback similarly warned that a flexible, multi-regulator approach "could stifle innovation" if companies lacked a single, authoritative steer on responsible deployment.

Part of this bottleneck sits within firms' own compliance teams. Their role - and career incentive - is to avoid regulatory missteps, so they default to extreme caution. When a new technology falls outside familiar control frameworks, the safest course often becomes a blanket refusal or a request for explicit regulatory sign-off. Sometimes, a clear, detailed rule provides exactly what firms need: certainty about where the guardrails sit, giving new entrants a common standard to build against.

The optimal regulatory approach for AI, therefore, is a **hybrid model** that combines principles-based flexibility with sufficient specificity to provide market certainty. This involves:

- **Core principles** that articulate desired outcomes (proportionality, accountability, resilience) without prescribing specific technical implementations.
- **Outcome-oriented benchmarks** that translate principles into measurable thresholds - such as maximum acceptable false positive rates for bias detection or minimum explainability requirements for high-stakes decisions.
- **Safe harbours and guidance** that provide clarity on acceptable practices without ossifying into rigid rules.
- **Regulatory sandboxes** that allow experimentation under supervision, generating evidence to refine both principles and specific guidance over time.

This hybrid model is best expressed as a **modular regulatory structure** operating across three layers: stable primary law, which sets objectives, duties, supervisory powers and the boundaries of delegation, and is deliberately hard to change; updatable secondary instruments - rules, classification schedules, thresholds and technical standards - revisable through defined, authorised procedures with consultation and publication requirements; and non-binding guidance, which moves fastest and carries interpretive rather than legal force. The distinction matters because flexibility and adaptivity are different properties: flexibility is discretion exercised within existing rules, while adaptivity is the institutional capacity to change the rules themselves through a lawful procedure. A framework can be flexible yet unable to adapt - and, as Section 3.1.1 shows, a framework whose adaptation depends on emergency primary legislation is not yet future-responsive. Modularity localises revision in the layers built for it.

4.2 Iterative Development and Regulatory Experimentation: Learning from AI Sandboxes and Policy Prototyping

The governance of complex, dynamic systems - from modern software development to urban planning - is best achieved through iterative processes of experimentation, feedback, and adaptation. The traditional "one-shot" or "waterfall" model of rulemaking, which aims to create a perfect

and comprehensive set of rules from the outset, is fundamentally unsuited to the inherent uncertainty and rapid evolution of artificial intelligence. An agile regulatory approach, by contrast, assumes that initial assumptions will be flawed and treats mistakes as essential learning opportunities. It does not aim for perfection on day one; instead, it builds in formal mechanisms to detect errors, correct course, and iterate.

Regulatory sandboxes, as discussed in Section 3.3.2, are the primary institutional tool for this iterative process. They create a structured feedback loop where regulators and innovators can learn together in a controlled, real-world environment. The insights gathered from these experiments - on the true capabilities of a technology, its emergent risks, and the effectiveness of various safeguards - should not be treated as one-off outcomes. For an adaptive framework to function, these learnings must be systematically collected, analysed, and used to inform the next iteration of the regulatory framework, creating a continuous cycle of improvement.

A related and complementary technique is **policy prototyping**. This experimental governance method involves the co-creation and testing of new regulatory approaches and policy instruments before they are enacted on a wide scale. It effectively applies the sandbox concept not just to a technology, but to the policy itself. A prominent example is the "Open Loop" programme. This multi-stakeholder initiative brought together regulators, technology companies, and civil society representatives to prototype and test different models for conducting AI impact assessments, providing invaluable empirical data on what works in practice.

For this iterative cycle to be effective, this agile approach to oversight system must establish formal **feedback loops** that translate experimental learnings into policy evolution. These mechanisms are essential for closing the gap between experimentation and implementation. They could include:

- Mandatory public reporting on the aggregated outcomes, key findings, and lessons learned from all national regulatory sandbox cohorts.
- The establishment of a dedicated regulatory unit (as discussed in Section 6.1.2) tasked with the specific mandate of analysing sandbox data and proposing concrete updates to official guidelines, outcome-oriented benchmarks, or technical standards.
- A legal requirement for regular, scheduled reviews of the overall regulatory framework (e.g., every two to three years) with the explicit goal of incorporating new evidence from sandboxes, real-world deployments, and international best practices.

Successfully embracing iterative development requires more than just launching new programmes; it demands a profound cultural shift within regulatory bodies themselves. The traditional regulatory mindset is one of "enforcement," where the primary function is to ensure compliance with a fixed set of rules, and success is measured by metrics like compliance rates and penalties issued. An agile regulator, however, must adopt a mindset of "stewardship," where the primary function is to guide the healthy and responsible development of a technological ecosystem. In this paradigm, success is measured by a more complex and nuanced set of indicators: the rate of safe innovation, the effectiveness of risk mitigation across the industry, and the level of public trust in the technology.

This shift has significant implications for the skills, processes, and performance metrics of regulatory agencies. It requires a new breed of regulator - one who is not only a legal expert but also a facilitator, a convener, and a strategic partner. These individuals and teams need skills in data analysis, experimental design, stakeholder management, and systems thinking. The internal processes of the agency must also transform, moving away from slow, sequential "waterfall" procedures toward more agile, iterative "sprint" cycles, mirroring the methodologies that have enabled the technology industry to innovate

so rapidly. Therefore, implementing a truly iterative regulatory model is not just a matter of launching sandboxes; it requires a deep, organisational transformation of the regulatory bodies themselves. This is the central challenge of building regulatory capacity, a topic that will be addressed in detail in Section 5.4.

4.3 Continuous Oversight and Dynamic Risk Management: Real-Time Monitoring and Automated Auditing

Traditional, static audits cannot adequately safeguard AI systems that evolve after deployment. Adaptive governance requires continuous oversight, combining automated monitoring, persistent testing, and dynamic risk evaluation.

Continuous Security Validation.

AI systems demand ongoing red teaming and automated testing to maintain trust over time. Techniques such as persistence testing help determine whether previously fixed vulnerabilities reappear after model updates or retraining. This continuous validation ensures that security controls remain effective as the AI adapts to new data, contexts, or operating environments.

Dynamic Risk Management.

A continuous AI red teaming approach integrates both automated scanners and expert human testing to detect new or regressed vulnerabilities in real time. When issues are discovered, they feed directly into risk management workflows, triggering remediation, review, and documentation updates.

Automated Auditing Infrastructure.

Emerging tools enable real-time auditing of model behaviour, data integrity, and access control through API hooks and monitoring agents. These systems generate live metrics and alerts tied to pre-defined benchmarks for fairness, safety, and performance.

Lifecycle Oversight.

Continuous validation transforms AI governance from a reactive process into a proactive one. By maintaining constant visibility over an AI system's operational state, organisations can demonstrate dynamic compliance, prevent drift-related risks, and uphold long-term accountability.

Legitimacy Safeguards for Continuous Oversight.

Continuous monitoring changes the relationship between supervisor and supervised, and it is not legitimate by default. An adaptive framework must therefore pair oversight powers with explicit safeguards, specified in the enabling instrument rather than left to practice. Monitoring requires a defined statutory basis and purpose limitation: telemetry, logs and model access may be demanded only for stated supervisory purposes, proportionate to the system's risk tier, and data collection should be minimised and aggregated wherever supervision permits. Commercially sensitive material - model weights, training-data details, security findings - requires protection through secure inspection environments, confidentiality obligations on supervisory staff and controlled disclosure procedures, so that assurance does not become a channel for trade-secret leakage. Due process applies to monitoring outcomes as it does to enforcement: operators are entitled to know the benchmarks against which they are monitored, to receive reasons when findings are adverse, to make representations before consequences attach, and to access review independent of the deciding supervisor (Section 5.4.1). Finally, oversight itself requires oversight: monitoring policies should be published, their use audited, and the supervisor accountable for proportionality - the same discipline the framework demands of the systems it governs.

4.4 Designing for Safety and Robustness: Ethical "Circuit Breakers," Rigorous Testing, and Verifiable Fail-Safes

Building safe and reliable AI requires more than implementing defensive features, it demands rigorous, adversarial testing to prove that those defences hold under real-world pressure.

Adversarial Stress Testing.

Testing plays a critical role in ensuring system robustness, with AI Red Teaming serving as a primary method for evaluation. Red teams deliberately attempt to provoke unsafe, biased, or unintended outputs by exploiting edge cases or adversarial inputs. These exercises reveal where existing safeguards may fail and where additional or strengthened controls are required.

Designing Ethical Circuit Breakers.

When testing reveals that an AI can produce harmful outputs or unsafe decisions, developers should implement "circuit breakers", automated mechanisms that halt operations, escalate alerts, or switch to safe modes. Examples include response throttling, human-in-the-loop validation, or isolation of compromised subsystems.

Fail-Safe Verification.

Every safety control should undergo empirical validation to confirm that it functions reliably under stress conditions. This includes simulating malicious prompts, data corruption, or model drift. The verification process must be repeatable, documented, and independently testable.

Safety by Design and Testing by Default.

Embedding adversarial testing and fail-safe design into the development lifecycle ensures that safety mechanisms evolve alongside the AI system. Rigorous, scenario-based testing is the most reliable available means of establishing that AI systems can withstand the pressures actually tested for while maintaining ethical and operational integrity; it is necessary evidence, not a guarantee of behaviour under conditions the tests did not cover.

4.5 Embedding Algorithmic Accountability: Transparency, Explainability, and Fairness by Design

Algorithmic accountability is not a post-hoc compliance exercise, it is an **architectural property** of a well-governed AI system. A system is accountable when every decision it makes can be inspected, explained, and justified in relation to its inputs, objectives, and constraints. This requires embedding transparency, explainability, and fairness directly into system design, not treating them as policy add-ons.

Design-Time Controls

Accountability begins at design. Models and pipelines must include:

- **Interpretability layers** (e.g. SHAP and LIME which suit classic machine learning far better than generative AI) to reveal which inputs drive predictions.
- **Decision logging** with immutable event trails for model inferences, thresholds, and override actions.
- **Feature attributions** stored alongside prediction metadata to support internal review.
- **Consented data usage** enforced through data-governance APIs that restrict access to approved purposes.

All fairness and explanation requirements should be captured as **non-functional controls** in design documentation, treated with the same priority as performance or reliability specifications.

Decision Lineage and Traceability

Every AI decision should be reconstructable end-to-end. Maintain input -> feature -> model -> policy -> action lineage using version-controlled artefacts: dataset hashes, feature pipelines, model versions, and configuration parameters. Such provenance enables technical audits, user appeals, and forensic analysis in the event of disputes or incidents. Lineage should align with the audit-trail principles defined in ISO/IEC 42001:2023 and NIST AI RMF "Govern/Manage"functions.

Meaningful Explanations

Explainability must match its audience:

- **Operators** require technical rationale to debug or validate outcomes.
- **Customers or citizens** need plain-language reasons for decisions that affect them, and clear recourse pathways.
- **Regulators and auditors** need reproducible evidence and explanation logs showing compliance with fairness and safety thresholds.

Human-override or escalation triggers should be embedded at decision points where confidence or explainability scores fall below governance-defined thresholds.

Fairness as a Hard Constraint

Fairness targets should be enforced as **model constraints**, not aspirational goals. When monitored metrics (e.g., disparate impact ratio, equalised odds) breach thresholds, deployment pipelines must trigger automatic **rollback or model-freeze**until remediation is verified. This converts ethical guidance into enforceable operational control, aligning with the "assurance by design" philosophy promoted in current EU AI Act and NIST frameworks.

Governance Reviews and Oversight

Accountability mechanisms must be visible at the top of the governance stack. Boards or AI-ethics committees should periodically review:

- transparency and interpretability reports;
- fairness and drift metrics over time;
- unresolved exceptions or user complaints;
- remediation and retraining status.

Document reviews, sign-offs, and decisions to form a verifiable audit trail that can withstand regulatory or legal scrutiny.

Framework Integration and Evidence of Compliance

Link all accountability artefacts, logs, lineage maps, fairness metrics, and review records - to a formal AI Management System (AIMS) under ISO/IEC 42001, and to the risk-management controls in ISO/IEC 23894 and NIST AI RMF. These mappings provide demonstrable compliance, harmonise terminology, and support external assurance or certification.

4.6 Multi-Stakeholder Collaboration: Engaging Industry, Academia, and Civil Society in Governance

The governance of artificial intelligence cannot be the exclusive domain of government, as AI systems are inherently socio-technical in nature. Their development, deployment, and societal impact are shaped by a complex interplay of technology, human behaviour, economic incentives, and social context. Consequently, effective and legitimate governance requires the active and continuous engagement of a diverse range of stakeholders. This multi-stakeholder approach ensures that regulatory frameworks are well-informed by technical expertise, grounded in empirical evidence, and aligned with broad societal values and the protection of fundamental rights.

A comprehensive governance model must create

formal and accessible channels for engaging with key stakeholder groups. These include:

- **Industry:** AI developers, deployers, and infrastructure providers who possess the deepest technical expertise and a practical understanding of the technology's capabilities, limitations, and commercial realities.
- **Academia:** Researchers from a wide range of disciplines, including computer science, law, ethics, economics, and the social sciences, who can provide independent analysis, critical evaluation, long-term foresight, and empirical evidence to inform policy.
- **Civil Society and Advocacy Groups:** Non-governmental organisations, consumer protection groups, and organisations representing the public interest, specific communities, and vulnerable populations. These groups play a crucial role in highlighting potential harms, advocating for rights-preserving safeguards, and ensuring that the voices of those most affected by AI are heard.
- **The General Public:** The end-users of AI systems and the citizens whose lives will be impacted by their deployment. Building public trust is essential for the successful and sustainable adoption of AI, and this requires transparency and meaningful opportunities for public input.

To be effective, collaboration must be institutionalised rather than ad hoc. An adaptive governance framework should incorporate several formal mechanisms for multi-stakeholder engagement:

- **Formal Advisory Boards:** The establishment of permanent, multi-stakeholder advisory councils or ethics committees tasked with providing ongoing expert advice to regulators and policymakers. This structure mirrors the successful multi-stakeholder composition of international bodies like GPAI.
- **Open Public Consultations:** The consistent use of transparent and accessible public consultation processes for all draft regulations and major policy initiatives, ensuring that a wide range of perspectives are solicited and considered before rules are finalised.
- **Co-Design in Regulatory Labs:** The explicit design of regulatory sandboxes and policy prototyping initiatives as platforms for formal co-creation. These labs should not just be for companies and regulators but should include structured participation from academic and civil society representatives to ensure a balanced and holistic approach to experimentation.
- **International Fora:** Active participation in and support for international multi-stakeholder alliances, such as GPAI and the UAE-WEF GRIP platform. These fora are essential for sharing best practices, learning from global experiences, and fostering the international coherence needed to govern a borderless technology.

Key Components of an Adaptive Governance Framework for AI

5.0 Key Components of an Adaptive Governance Framework for AI

5.1 A Tiered, Risk-Based Approach: Differentiating Regulatory Scrutiny Based on AI Capability and Potential Impact (Informed by Models like the EU AI Act)

A tiered, risk-based approach is the framework's central proportionality device: it concentrates supervisory attention and compliance cost where the potential for harm is greatest, and leaves low-risk innovation substantially free. The rationale is both economic and institutional. Uniform obligations either under-protect against the highest-impact systems or over-burden the lowest; and no supervisor has the capacity to scrutinise every deployment, so classification is also a resource-allocation mechanism. The EU AI Act is the most developed legislated model of risk differentiation and informs the design here, with two deliberate departures explained by the analysis in Section 3.1.1: the tiers below are regulatory design categories defined by this framework for its own purposes, not a restatement of any other statute's operative structure; and classification is built from the outset to be revised.

The framework distinguishes four categories. **Prohibited uses** form a narrow, exhaustively defined list of applications whose harm is judged categorically unacceptable; the list is set in primary law and amendable only through it. **High-impact systems** are those whose failure or misuse can significantly affect safety, fundamental rights, critical infrastructure or market integrity; they carry the framework's core obligations - risk management, data governance, documentation, human oversight and the continuous assurance regime of Section 5.2 - evidenced before deployment and maintained through the lifecycle. **Transparency-tier systems** interact with people or generate synthetic content in ways that warrant disclosure and labelling duties without full assurance obligations. **Baseline systems** attract only generally applicable law and voluntary good practice.

Classification is determined by criteria assessed in combination rather than by any single attribute: the intended use and its deployment context; the degree of autonomy and the reality - not merely the design - of human oversight; the scale of deployment and the population affected; the severity and reversibility of plausible harm; the sensitivity of the data processed; the system's interconnection with other critical systems; and the capability class of the underlying model, including general-purpose capability and tool access. Context dominates capability at the margin: the same model may be baseline in one deployment and high-impact in another, which is why the framework classifies deployments rather than technologies (Section 1.5).

What distinguishes this component from static tiering is that classification is an output of the capacity cycle, not a one-time gate. Sensing inputs - incident reports, complaint patterns, red-team findings, post-market monitoring and notified model updates - feed defined triggers for reclassification review, conducted by the classification authority named in the enabling instrument under a published procedure with reasons, effective dates and transition periods for reclassified systems (Sections 5.4.1 and 6.1.2). Providers bear a duty to notify capability changes material to classification, mirroring the change-governance obligations of Section 5.2. Misclassification is treated as an expected event to be detected and corrected, not an anomaly: the periodic review of the classification schedule, informed by the evaluation stage

of the cycle, is where the framework's future-responsiveness becomes operational rather than rhetorical.

5.2 Mandating Continuous Assurance for High-Impact and Frontier AI Models

5.2.1 Integrating Evolving Red Teaming Standards and Methodologies (a proposed AI Red Team Handbook for ADGM)

This section sets out the design of a **proposed** AI Red Team Handbook for ADGM. No such instrument has been published: the handbook is presented as a proposal for consideration by the Registration Authority, to be developed under the enabling instrument discussed in Section 5.4.1, and the material below describes what it should contain.

AI red teaming must extend beyond classic penetration testing to cover data, model, environment, and governance layers, and evolve as systems and threats do.

- **Methodology by design.** Define objectives, attacker models, and safety constraints alongside functional requirements; treat red teaming as a first-class lifecycle activity tied to change management gates.
- **Scenario libraries.** Include poisoning (label flipping, backdoors, data pipeline tampering), extraction (weights/API semantics), evasion (adversarial inputs), prompt/goal injection (including indirect injections via retrieved or tool output), reward hacking, safety-bypass, supply-chain compromise, and cyber-physical manipulation. Use MITRE ATLAS to structure TTPs.
- **Agentic and Tool-use coverage.** For systems with tools, memory, or planners, test orchestration flaws (unsafe tool calling, permission escalation), state-carryover risks, and long-horizon goal mis-generalisation. See contemporary agentic red-teaming guidance for concrete test patterns.
- **Operational realism.** Exercise in sandboxed but production-faithful environments with representative tools, telemetry, and human-overrides; capture MTTD/MTTR for unsafe actions and "near-miss" analytics to quantify control effectiveness.
- **Structured reporting.** Standardize findings: assumptions, exploit path, evidence (inputs/outputs/telemetry), impact grading (safety, compliance, financial), recommended mitigations, and re-test results; link to risk registers and change tickets.
- **Continuous cadence.** Trigger re-tests on model updates, distribution shifts, policy changes, or vendor integrations; ingest incidents and AVID entries into new scenarios so learning compounds.
- **Regulatory context.** Where ADGM/sectoral guidance mandates red-team style testing, map your programme to those expectations and keep artefacts regulator-ready.

5.2.2 Threat Modelling as a Foundation for Continuous Assurance

While AI red teaming provides empirical testing of system robustness, effective assurance begins earlier, with structured threat modelling. Threat modelling transforms abstract AI risks into concrete, repeatable controls throughout the model lifecycle. It enables organisations and regulators to identify what must be defended, from whom, and which mitigations provide the greatest risk reduction. The following framework outlines how adaptive governance can embed threat modelling as a mandatory component of AI system assurance,

ensuring alignment with evolving standards such as MITRE ATLAS and ISO/IEC 42001.

5.2.2.1 The Case for Threat Modelling in Adaptive AI Governance

AI systems have become simultaneously **critical and fragile**.

Production AI systems are increasingly mission-critical to the organisations that deploy them, while dedicated AI-specific threat management remains immature in many of those organisations. At the same time, integration with third-party generative services, ChatGPT, Copilot, Gemini, and similar, has expanded the attack surface far beyond traditional software boundaries.

This convergence of dependence and exposure defines the governance imperative for threat modelling:

- **Bridging the "pacing gap" between innovation and control.** Threat modelling transforms rapid deployment pressure into a disciplined, repeatable cycle of risk assessment aligned with adaptive regulation.
- **Operationalising assurance principles.** It links high-level mandates, such as continuous oversight and dynamic risk management to concrete engineering artefacts that can be examined by regulators or auditors.
- **Embedding security-by-design.** Traditional frameworks focus on code, identity, and network threats. Frontier AI introduces new attack planes: model artefacts, multimodal inputs, autonomous agents, and AI supply chains. Without structured threat modelling, these remain invisible to policy oversight.
- **Providing measurable assurance evidence.** A maintained threat model offers regulators traceability, what threats were considered, what mitigations applied, and how effectiveness is validated, satisfying lifecycle documentation under ISO/IEC 42001 and complementing AI red-team reporting.

In adaptive governance, threat modelling therefore functions as the first line of continuous assurance, guiding where red teaming, auditing, and monitoring should focus, and enabling proportional, risk-based oversight.

5.2.2.2 Lifecycle Blueprint: From Design to Continuous Oversight

An adaptive threat-modelling programme must mirror the entire AI lifecycle, ensuring that security assumptions remain valid as models evolve, retrain, or integrate with new tools. The following eight-stage blueprint provides a structured workflow that regulators and developers can jointly adopt:

- **Define mission and criticality.** Determine the business or societal function the AI supports and what constitutes unacceptable harm. This anchors risk prioritisation in real-world impact.
- **Inventory assets and trust boundaries.** Maintain a living catalogue of models, datasets, prompts, indexes, connectors, agents, and third-party integrations. Mark where untrusted data or commands can enter the system.
- **Decompose architecture.** Map data and action flows across model-only, RAG, agentic, and workflow-integrated architectures, highlighting interfaces with external content and tools.
- **Enumerate threats.** Identify attack vectors specific to AI prompt and multimodal injection, data poisoning, model theft, serialisation exploits, agentic misuse, GPU and MLOps vulnerabilities and trace how each could propagate through the system.
- **Prioritise and map to frameworks.**

Score likelihood and impact; align mitigations with recognised taxonomies such as MITRE ATLAS and the OWASP LLM Top-10 to enable interoperability with global standards.

- **Design layered controls.**
  - Pre-deployment: prefer safe serialisation formats (e.g., SafeTensors), scan models for tampering or backdoors, sandbox deserialisation, document provenance, and verify datasets for poisoning.
  - Runtime: enforce retrieval-side sanitisation, cross-modal filters, scoped API tokens, and human-in-the-loop approval for high-impact actions.
  - Supply chain: require signed artefacts, hardened MLOps consoles, and regular GPU/container patching.
- **Validate like an attacker.**
  Integrate adversarial ML testing, LLM and AI Agents red teaming before release and after each major update to confirm that mitigations remain effective.
- **Monitor and respond.**
  Treat AI-specific incidents prompt injections, tool poisoning, RAG poisoning, capability misuse as first-class operational events. Log prompts and tool calls, detect anomaly patterns, quarantine compromised indexes, and define rollback procedures.

This cyclical model converts governance principles into a living control system, allowing both developers and regulators to trace risk decisions across time and version changes, an essential requirement for frontier AI oversight.

5.2.2.3 Workflow-Specific Threat Maps and Governance Patterns

Different AI deployment patterns exhibit distinct threat surfaces and require tailored governance responses. Embedding threat modelling within sectoral or workflow-specific assurance regimes allows regulators to apply proportional scrutiny.

a. Retrieval-Augmented Generation (RAG) Assistants

- Primary threats: poisoned source documents, injection during ingestion, retrieval of hostile snippets, data leakage through citations.
- Controls: approved-source curation, checksum-based indexing, retrieval allowlists, context-length caps, output validators, and “sensitive-topic” circuit breakers.
- Governance linkage: mandatory content provenance verification before public deployment.

b. Side-Panel Copilots (e-mail, documents, slides)

- Threats: indirect prompt injections in user content, over-privileged scopes, unsanctioned send actions.
- Controls: HTML/OOXML sanitisation, “read-only-by-default” policies, explicit user consent for outbound actions, DLP and policy guards, full audit trails.
- Governance linkage: consent and accountability requirements under data-protection regulation.

c. Agentic AI (planners and tool-using agents)

- Threats: unsafe or escalated tool invocation, goal hijacking, environment manipulation.
- Controls: per-tool allowlists, least-privilege credentials, deterministic wrappers for shell/database/payment functions, execution sandboxes, and telemetry on tool calls.
- Governance linkage: certification of tool-safety policies and autonomous-action thresholds.

d. Code Assistants and ChatOps Integrations

- Threats: poisoned repositories or issue comments, secrets echoing in suggestions, triggered CI actions.
- Controls: secret scanning, signed CI actions, provenance metadata for suggestions, enforced human review gates.
- Governance linkage: secure-software-development obligations and change-management requirements.

e. MLOps Pipelines and Model Registries

- Threats: malicious or tampered model artefacts, unsafe loaders, exposed consoles, credential leakage.
- Controls: model signing and scanning, isolation of conversion processes, SafeTensors enforcement, token rotation, closed management interfaces, and honeypot monitoring.
- Governance linkage: supply-chain assurance and mandatory incident-reporting channels to AI-specific vulnerability databases (e.g., AVID).

Across all workflows, the guiding pattern is consistent: "ship with a model-aware gate."

No high-impact AI system should be deployed without an up-to-date threat model, recent red-team results, and a rollback plan. Adaptive regulators can require these artefacts as part of ongoing assurance submissions, ensuring that the pace of innovation is matched by the maturity of defence.

5.2.3 Establishing AI Bias Audit Protocols and Fairness Metrics

Bias is not only an ethical issue but an operational, regulatory, and trust-governance risk. AI systems that influence credit, hiring, healthcare, or public services are already under scrutiny in multiple jurisdictions (e.g., EU AI Act Chapter III high-risk systems; US FTC and EEOC guidance). Audits must therefore be lifecycle-based, evidence-driven, and context-specific, aligning fairness evaluation with measurable impact on affected groups.

Scope and Inventory

Catalogue all models whose outputs affect access to opportunities, pricing, or benefits. Identify protected attributes (race, gender, age, geography, disability, socio-economic status) and any proxy variables that may indirectly encode them. Each model entry should include: purpose, data sources, decision domain, and regulatory classification. Maintain linkage to the organisation's AI risk register.

Data Audit

Quantify representativeness, label accuracy, sampling balance, and missingness. Compute correlations between input features and protected classes to reveal hidden proxies. Document known historical bias, imbalanced labels, or data-collection artefacts. Where feasible, apply dataset shift detection and bias heat-maps to visualise exposure. Auditors should confirm that lineage and versioning for datasets are preserved, enabling replication.

Metric Selection and Thresholds

Select fairness metrics aligned to potential harm types and model function:

- **Demographic parity** - equal positive-prediction rate across groups.
- **Equalised odds** - equal true-/false-positive rates across groups.
- **Predictive parity & calibration by group** - probabilistic fairness for risk scores.
- **Disparate-impact ratio** - including the four-fifths (0.8) screening heuristic drawn from the US Uniform Guidelines on Employee Selection

Procedures (29 CFR §1607.4(D)); an adverse-impact screen for that specific context, not a universal AI fairness standard, and to be paired with context-specific metrics (Section 3.3.3).

Define tolerance thresholds and escalation paths approved by governance boards. Threshold exceedance must trigger remediation or model suspension depending on criticality.

Testing and Simulation

Perform pre-deployment and recurring tests under varied data and context conditions:

- Attribute-perturbation tests (alter sensitive attributes and observe outcome changes), which are a useful sensitivity check but approximate rather than implement causal counterfactual fairness in the sense of Kusner and others (2017).
- Stress tests for distributional drift, under-represented cohorts, and edge cases.
- Adversarial bias-amplification testing, inject manipulated data to probe model sensitivity.

Automate fairness regression testing so new model iterations are validated before promotion.

Mitigation and Monitoring

Apply fairness interventions appropriate to stage and model type:

- **Pre-processing** - re-sampling, re-weighting, synthetic data augmentation.
- **In-processing** - fairness-aware loss constraints, adversarial debiasing.
- **Post-processing** - threshold adjustment, output calibration, or rule-based correction.

Validate improvements on hold-out datasets and monitor real-world subgroup outcomes. Deploy drift detection for subgroup-level metrics (e.g., false-negative disparities). Trigger alerts on threshold breaches and route to bias-governance escalation.

Assurance Evidence and Vendor Accountability

Maintain complete audit workpapers:

- datasets examined and versions;
- metrics computed and thresholds;
- exceptions, remediation plans, and re-test results;
- governance sign-offs and timestamps.

Third-party or open-weight model vendors must provide equivalent fairness attestations and redress mechanisms under contractual clauses.

Align documentation with recognised frameworks such as NIST AI Risk Management Framework (Govern/Map/Measure/Manage) and ISO/IEC 42001 AI Management System controls for fairness auditing.

Bias auditing is a continuous control loop, measure, stress, mitigate, and verify. Treat fairness metrics as operational SLOs, not research statistics. When fairness is governed as part of system assurance rather than ethics alone, AI becomes demonstrably accountable and legally defensible.

5.2.4 Compliance Assurance Through Integrated AI GRC and Privacy Frameworks

An effective AI assurance regime cannot rely solely on technical assessments such as red teaming or threat modelling. To align with evolving global regulatory expectations (e.g., EU AI Act, ISO/IEC 42001, NIST AI RMF, GDPR, and OECD AI Principles), organisations must embed **AI Governance, Risk, and Compliance (GRC)** and **Data Protection** principles across the entire AI lifecycle - from design to decommissioning.

Drawing on Hacken's field-tested frameworks for LLM and AI system security, this contribution advocates for the integration of **AI GRC and Data Privacy Compliance as structured control domains**, ensuring not only that technical mitigations exist,

but that they are backed by enforceable policies, risk oversight, and demonstrable regulatory conformity.

5.2.4.1 Key AI GRC Dimensions

Governance: Assign clear ownership over AI systems, with defined roles for AI risk officers, red team leads, DPOs, and model maintainers. Governance structures should define escalation paths, sign-off checkpoints, and accountability responsibilities aligned with each lifecycle phase.

Risk Management: Operationalise threat modelling outputs and red teaming reports into risk registers, treatment plans, and compensating controls. This ensures that assurance activities produce measurable improvements in system security posture.

Compliance Register: Maintain structured evidence that links controls to legal and industry requirements - such as pre-deployment testing, red-team result archives, audit logs, and secure model update records - offering traceability to regulators and auditors.

5.2.4.2 Embedding Data Privacy and AI-Specific DPIAs

(Contributors: Dmytro Yasmanovych, Hacken)

Data privacy compliance must be treated as a parallel assurance track, especially for generative and foundation models that process or generate personal data (PII), biometric data, or sensitive content.

Key privacy controls to enforce across the AI pipeline:

Lawful Basis and Consent Management

- Clearly define the legal basis (e.g. legitimate interest, consent) for AI data collection and processing activities.
- Where consent is the applicable legal basis, implement purpose-specific consent for high-risk uses, particularly in healthcare, education, or financial AI applications; consent is not the default basis for high-risk AI, and the legal basis must be identified for each processing purpose before any consent mechanism is designed.

Data Minimisation & Purpose Limitation

- Avoid overcollection of training and inference data.
- Clearly define the boundaries between operational and experimental data streams in multimodal and RAG systems.

AI-Specific DPIAs (Data Protection Impact Assessments)

- Conduct DPIAs at the level of each high-risk processing operation, as data-protection law requires, and - as the governance measure proposed here rather than a description of current law - maintain and update one per high-risk model, using an adapted template that includes:
  - Model type and purpose
  - Types of data processed (including embeddings, prompts, logs)
  - Intended use cases and foreseeable misuse
  - Data retention strategy and access restrictions
  - Risk mitigation strategies (technical and organisational)
- Link DPIAs to threat modelling and red teaming results to create an integrated assurance narrative.

Output Monitoring for Data Leakage

- Monitor model outputs for unintended memorisation, data leakage, or PII echoes - especially when fine-tuning on user data or

third-party documents.

- Enforce retrieval-side data sanitisation, inference-level output validators, and human-in-the-loop review where privacy risks are elevated.

5.2.4.3 Leveraging Global Standards for Privacy and AI Alignment

Use of internationally recognised privacy and AI frameworks ensures interoperability with future regulation and simplifies third-party assurance. Examples include:

- **ISO/IEC 27701:2025**: Privacy information management system requirements (second edition, October 2025; now a standalone management-system standard rather than an extension of ISO/IEC 27001)
- **ISO/IEC 42001**: AI Management System requirements
- **NIST AI RMF (2023)**: Risk management and governance overlays
- **EU AI Act & GDPR**: Risk-tier classification and personal data governance
- **OECD AI Principles**: Transparency, accountability, human-centricity

Incorporating GRC and Data Privacy into AI assurance closes the accountability loop - ensuring that technical safeguards are matched with governance, documentation, and lawful processing controls. Adaptive regulators should require the **ongoing maintenance of an AI Risk Register, a Privacy Impact Register, and a Compliance Register**, covering both internal operations and third-party integrations.

These living artefacts enable continuous oversight and lifecycle alignment, ensuring that security, trust, and legality scale with innovation.

5.3 Collaborative Foresight and Horizon Scanning: Public-Private Partnerships for AI Risk Analysis and Anticipatory Governance

5.3.1 Learning from Initiatives: The UAE's Use of AI in Legislative Processes (Regulatory Intelligence Office) as an example of tech-enabled foresight

The United Arab Emirates is pioneering a novel approach to regulation that moves beyond simply governing AI to actively using AI to make the regulatory process itself more intelligent, agile, and forward-looking. In what the UAE Government presented as a world first, the UAE Cabinet approved the launch of the **Regulatory Intelligence Ecosystem** and the establishment of the **Regulatory Intelligence Office** in April 2025. This initiative represents a powerful case study in tech-enabled foresight and provides a model for how governments can build institutional capacity for anticipatory governance.

The core function of this AI-powered ecosystem is to create a "living, responsive" legislative framework that can adapt in near real-time to technological and societal changes. Its key capabilities include:

- **Comprehensive Legislative Mapping:** The system creates a unified digital map of the entire legislative landscape, connecting all federal and local laws with associated judicial rulings, executive procedures, and public service data.
- **Real-Time Impact Analysis:** By leveraging large-scale data analytics, the ecosystem can track the daily impact of laws on the economy and society across more than 20

sectors, providing an empirical basis for policy evaluation.

- **Automated Gap and Conflict Detection:** The system is designed to autonomously identify outdated, redundant, or conflicting regulations among the more than 100 federal and local laws currently in force.
- **Proactive Policy Recommendations:** The platform uses AI to suggest legislative updates and amendments based on its data analysis, emerging trends, and continuous benchmarking against international best practices sourced from global research centres.

The anticipated outcomes of this initiative are transformative. The UAE government projects that the system will accelerate the entire legislative cycle - from initial research and drafting to final evaluation - by as much as 70%. This dramatic increase in efficiency allows the government to shift from a traditionally reactive policymaking posture to one that is anticipatory and proactive, directly addressing the core challenge of the pacing problem. The Regulatory Intelligence Office thus provides a powerful lesson for adaptive governance: governments can and should leverage the very technologies they seek to regulate to enhance their own foresight and agility.

The UAE's initiative represents the ultimate feedback loop within an adaptive governance system. A core principle of a more dynamic regulatory posture is the necessity of formal feedback mechanisms to learn from real-world outcomes and adjust policies accordingly. Traditionally, this process is slow, manual, and often subject to political influence, relying on periodic legislative reviews or commissioned academic studies. The UAE's approach automates and accelerates this feedback loop. It creates a direct, data-driven connection between the law as written and the law as experienced by analysing real-time data on economic and social impacts. This transforms legislation from a static text into a dynamic system that can be continuously monitored, measured, and optimised. It enables regulators to ask, on an ongoing basis, "Is this law achieving its intended outcome? If not, what data-driven adjustments can be made?" This constitutes a paradigm shift in governance, creating a model for a truly evidence-based, agile, and self-correcting regulatory state. It positions AI not just as an object of governance, but as a critical tool for better governance itself, potentially creating a new class of "Regulatory Technology" that could be exported globally.

Regulating with AI, however, imports the same risks the framework governs elsewhere, and a regulator that uses AI must hold itself to the standards it sets. Four disciplines follow. Decision support must be designed against **automation bias**: AI-generated analysis informs, but does not substitute for, accountable human judgement, and material decisions should record where AI input was relied upon. Regulatory AI must be **auditable**: versioned, logged and open to the same decision-lineage reconstruction required of supervised firms (Section 4.5). Outcomes influenced by AI must remain **contestable**, with affected parties able to challenge both the decision and the evidential role the system played in it. And accountability remains **human and institutional**: the deploying authority answers for the decision as if no AI had been involved. Embedding these disciplines in initiatives such as the Regulatory Intelligence Ecosystem is what separates tech-enabled foresight from an unexamined automation of public power.

5.3.2 The Role of Research Institutions (e.g., ADGM Academy) in Convening Stakeholders and Informing Policy

Research institutions occupy a distinctive position in an adaptive governance ecosystem: they can convene actors who cannot easily convene each other. A regulator that consults industry bilaterally invites capture concerns; industry actors that

coordinate directly raise competition sensitivities; academia alone lacks the practice context. A research centre attached to, but editorially distinct from, a regulatory authority - the ADGM Academy Research Centre is the model contemplated here - can host that triangle: structured dialogue between supervisors, firms, academics and civil society, conducted on the record and directed at published outputs rather than private positioning.

Three functions matter most. The first is evidence production: empirical work on regulatory performance, sandbox outcomes and emerging risk that individual authorities rarely resource internally, feeding the sensing and evaluation stages of the capacity cycle. The second is capability building: training regulators and market participants - the continuous professional development infrastructure that Section 5.4 identifies as a prerequisite for credible oversight. The third is policy experimentation: neutral ground for the policy prototyping of Section 4.2, where draft guidance and assessment templates can be tested with practitioners before formal consultation.

Editorial independence as a condition of the model. The value of a regulator-associated research centre rests on whether its findings can be trusted not to be the regulator's position in academic dress. The proximity that gives such a centre its convening power is also its principal vulnerability: readers will reasonably ask whether conclusions were shaped to suit the parent authority, and the answer has to be visible in the governance rather than asserted in a foreword. Four safeguards follow. First, an editorial board with a majority of members independent of the regulatory authority, holding the final decision on what is published. Second, published commissioning criteria, so that the choice of topics is open to scrutiny and cannot be read as agenda-setting by the authority. Third, disclosure of every author's institutional affiliation and actual role in each output; the contributions statement at the front of this paper is offered as the form that disclosure should take. Fourth, a clear line between two kinds of review. An authority may properly check statements about its own mandate, instruments and institutional facts for accuracy, and a centre publishing under an authority's brand may properly clear that use of the brand; neither is an approval right over analytical conclusions or recommendations, and a centre that grants one has ceased to be a research function. Where a centre cannot yet meet these conditions, the honest course is to say so in the output rather than to imply an independence the governance does not provide.

Partnerships extend the model's reach. Joint programmes with universities anchor the work academically - the collaboration with Abu Dhabi University reflected in this paper is an instance - while links to think tanks and international networks such as GPAI import comparative evidence. The output of this convening function is precisely the kind of document this paper aims to be: practitioner-grounded, academically reviewed and directed at implementable design.

5.4 Building Regulatory Capacity: Developing Technical Expertise and Agile Processes within Regulatory Bodies

A primary and persistent obstacle to the effective regulation of artificial intelligence is the significant expertise gap that often exists within government and regulatory bodies. Regulators, who are typically trained in law, public policy, or economics, frequently lack the deep, specialised technical knowledge of AI systems, data science, and cybersecurity required to conduct meaningful oversight, critically assess complex risks, and engage with industry experts on an equal footing. Closing this capacity gap is a prerequisite for implementing any credible adaptive governance framework.

To effectively govern AI, regulatory bodies must actively cultivate a new and interdisciplinary set of skills among their staff. This is not simply a matter of hiring a few data scientists; it requires a systemic

effort to build a broad base of expertise. Essential skills for the modern regulator include:

- **AI Literacy:** A foundational, conceptual understanding of how different types of AI models (e.g., machine learning, deep learning, generative AI) function, their core capabilities, and their inherent limitations.
- **Data Science and Statistics:** The ability to critically evaluate the quality and representativeness of datasets used to train AI systems, assess the statistical validity of performance metrics, and understand the principles of robust data governance to identify and mitigate potential biases.
- **Cybersecurity and AI Safety:** Specialised expertise in AI-specific vulnerabilities, such as data poisoning, model evasion, prompt injection, and the principles of secure and robust system design.
- **AI Red Teaming:** A practical knowledge of adversarial testing techniques and methodologies to understand how AI systems can fail under stress or when subjected to malicious attacks.
- **Legal-Tech Integration:** The crucial ability to bridge the gap between technical and legal domains, translating complex technical issues into clear, enforceable regulatory language and legal requirements.

Governments can and should adopt a multi-pronged strategy to build this essential capacity. Key approaches include:

- **Establishing Specialised AI Competence Centres:** Creating dedicated, centralised units or "centres of excellence" with a mandate to serve as a shared resource, providing deep technical expertise and analysis to all other government agencies and sectoral regulators (as will be discussed further in Section 6.1.2).
- **Talent Exchange and Fellowship Programs:** Implementing structured programmes that facilitate secondments and fellowships, allowing for the temporary exchange of personnel between regulatory bodies, leading technology companies, and academic research institutions. This fosters a crucial cross-pollination of expertise and perspectives.
- **Investing in Continuous and Advanced Training:** Developing and mandating continuous professional development programmes for all regulatory staff to ensure they remain current with the latest technological advancements and regulatory best practices.
- **Adopting Agile Internal Processes:** Restructuring the internal processes of regulatory agencies to be more agile and iterative. This involves breaking down large, multi-year rulemaking efforts into smaller, manageable "sprints" that allow for faster learning, quicker delivery of guidance, and more rapid adaptation to new information.
- **Leveraging Strategic Procurement:** Using the government's own procurement of AI systems as a powerful lever to build internal knowledge. Procurement contracts for AI systems should require vendors to provide extensive technical documentation, access to test environments, and comprehensive training for government staff as a condition of the contract.

5.4.1 An Enforcement Architecture for an Adaptive Framework

An adaptive framework without an enforcement architecture risks amounting to guidance by another name. The design criterion is therefore explicit: every binding duty requires an identifiable duty-holder, a named supervisor with adequate powers, evidential access, defined sanctions and independent review. This subsection specifies

that institutional minimum, using ADGM as the illustrative mapping. The mapping is a design proposal whose operationalisation would require an enabling instrument adopted through the ordinary legislative and policy process; it is not a description of current mandates.

Supervisory allocation. Each obligation in the framework must have a named supervisor before it takes effect. In the ADGM illustration, the proposed allocation follows existing institutional competence: AI applications in non-financial registered businesses would sit with the Registration Authority, whose current mandate covers registration, licensing and the enforcement of ADGM commercial legislation and would require extension by an enabling instrument to encompass AI-specific supervision; AI used in regulated financial services engages the FSRA within its regulated-activity perimeter; personal-data dimensions fall to the Office of Data Protection under the Data Protection Regulations 2021; and national-security and critical-infrastructure competence remains at UAE federal level (a layer consolidated on 14 June 2026, when the UAE Cabinet approved the establishment of the Artificial Intelligence and Data Authority as the single national body for data, artificial intelligence and digital government, absorbing the federal AI Office, the digital government sector of the TDRA and the UAE Data Office; UAE Cabinet announcement, 14 June 2026 - the constitutive instrument and commencement provision had not been published at the time of writing). The general principle is allocation by existing mandate rather than the creation of a monolithic AI regulator: AI risks manifest through activities that existing supervisors already oversee. Allocation alone does not resolve boundary cases - a single AI system can engage financial-conduct, data-protection and commercial rules simultaneously - so the coordination mechanism (Section 6.1.2's competence centre) must itself be given legal footing: a lead-supervisor rule, an information-sharing gateway and a defined conflict procedure are identified here as design requirements for the enabling instrument.

Complaint intake and triage. A single front door for AI-related complaints - operating as a no-wrong-door rule that routes each matter to the competent supervisor - serves both the public and the sensing stage of the future-responsiveness cycle. Complaint patterns are one early signal, to be triangulated with incident reports, supervisory findings and audit results, that a system's risk tier is mis-set or that guidance requires revision. Intake standards should specify acknowledgement and triage timelines and feed anonymised complaint statistics into the framework's evaluation stage.

Graduated sanctions. Enforcement should ascend a defined ladder: supervisory dialogue and remediation plans; binding conditions on deployment; financial penalties scaled to the gravity of harm and the operator's size; suspension or prohibition of the system; and, where the operator holds an authorisation, its revocation. The ladder must permit immediate escalation for urgent or serious harm, and each step requires a defined breach, a decision-maker and a procedure in the enabling instrument. Graduation is not leniency - it is what makes proportionate enforcement credible: supervisors will use instruments whose consequences are proportionate to the breach, and operators obtain the legal certainty needed to organise compliance.

Liability allocation. Where harm arises from an AI system assembled from a foundation model, third-party components and a deployer's configuration, responsibility must be traceable along the value chain. For regulatory obligations, the framework allocates duties to the party with control over the relevant risk - the provider for model-level properties, the deployer for context of use - with documentation duties (Sections 3.3.1, 5.2) supplying the evidentiary trail. Civil liability is a distinct question on which no settled international position yet exists: the EU now treats software and AI as products under the 2024 Product Liability Directive, its separate AI Liability Directive proposal was withdrawn, and national fault, causation and contract rules remain diverse. The framework

therefore treats civil liability - including insurance-based approaches to autonomous-system liability - as a priority for its first scheduled review rather than a matter it can presently harmonise.

Due process and review. Enforcement decisions require reasoned notice, a right to make representations, and access to review independent of the deciding supervisor - distinguishing internal reconsideration, any merits appeal the enabling instrument provides, and judicial review before the ADGM Courts. In an adaptive regime this is not procedural decoration: because obligations can change between authorisation and enforcement, legitimacy depends on non-retroactivity - operators judged against the version of the rules in force at the relevant time, supported by a published version register, effective dates and transition periods for reclassified systems.

5.5 Sector-Specific Considerations for Adaptive Governance (Illustrative Examples from UAE Priority Sectors)

5.5.1 Energy: Ensuring Safety, Reliability, and Cybersecurity in AI-Driven Energy Systems

The energy sector is rapidly integrating artificial intelligence to modernise infrastructure, improve efficiency, and support the critical global transition to renewable energy sources. Key applications being deployed in the UAE and globally are transforming how energy is generated, managed, and distributed. These include:

Grid optimisation, where AI is used to balance supply and demand in real-time, predict load, and prevent outages, a function that is essential for managing the inherent variability of renewable sources like solar and wind.

Predictive maintenance is another critical use case, where AI analyses vast streams of sensor data from infrastructure like pipelines, turbines, and transformers to predict equipment failures before they occur, thereby reducing costly downtime and enhancing operational safety. AI is also crucial for

Renewable energy forecasting, using sophisticated models to more accurately predict energy generation from solar and wind farms based on complex weather patterns and historical data.

However, the integration of AI into the heart of critical energy infrastructure introduces a new class of significant and high-stakes risks.

Cybersecurity threats are paramount; as energy systems become more digitised and interconnected, their attack surface expands, making them more vulnerable to sophisticated cyberattacks. AI can be used by malicious actors to orchestrate these attacks, and the AI-controlled systems themselves can become primary targets. The high degree of interconnection in an AI-optimised grid creates the risk of

Cascading failures, where a single failure - whether caused by a technical glitch, a cyberattack, or an unexpected emergent behaviour from the AI - could trigger a rapid, widespread blackout with severe economic and societal consequences. Furthermore, a failure in an AI system managing a physical facility like an oil and gas plant or a power station could lead to direct **physical safety and environmental harm**. Finally, because the performance of AI models is entirely dependent on the integrity of their input data, **data poisoning attacks** that target energy data could lead an AI system to make dangerously incorrect decisions about grid management or infrastructure control.

Given these heightened risks, an adaptive governance framework for AI in the energy sector must include specific, robust measures:

- **Mandatory AI Red Teaming:** For any AI system that exerts control over critical energy

infrastructure, rigorous and independent AI red teaming should be mandated. This adversarial testing is necessary to proactively identify vulnerabilities and critical failure modes under simulated attack conditions.

- **Robust Fail-Safes and Human Oversight:** All such systems must be designed with verifiable "circuit breakers" and robust fail-safe mechanisms that allow for immediate human intervention or an autonomous, safe shutdown in the event of anomalous or dangerous behaviour.
- **Supply Chain Security:** Strict, verifiable standards for data and model supply chain security must be enforced to protect against data poisoning and other upstream attacks that could compromise the integrity of the AI system.
- **Development of Sector-Specific Standards:** Regulators should work with industry stakeholders to develop and mandate sector-specific standards for the reliability, robustness, and cybersecurity of AI systems used in energy, potentially using regulatory sandboxes focused on energy applications to test and refine these standards.

5.5.2 Logistics & Transportation: Governance for AI in Autonomous Systems, Supply Chain Optimisation, and Smart Infrastructure

AI in logistics blends cyber and physical risk: autonomous movement of goods and people, algorithmic scheduling, and sensor-rich infrastructure. Governance must anticipate cascading effects across carriers, depots, ports, roads, rails, and energy.

Systems interdependence and safety.

Autonomous vehicles, drones, yard robots, and warehouse systems interact with humans and critical infrastructure. Objective functions (cost/throughput) must be constrained by safety, regulatory, and resilience guardrails. Reward-hacking and sensor spoofing are material risks.

Data provenance and integrity.

Routing, ETA, and maintenance decisions depend on heterogeneous data (telematics, EDI, IoT, geospatial, weather). Governance should enforce lineage, integrity checks, cryptographic authentication where feasible, and segregation of duties for data providers.

Optimisation versus resilience.

Single-point efficiency (e.g., just-in-time, single-sourcing) can create fragility. Boards should mandate stress tests (hub outages, strikes, cyber events, capacity shocks) and require policy-level buffers (inventory, multi-routing) encoded into optimisation constraints.

Explainability and traceability.

Operational staff and regulators need decision lineage for re-routes, prioritisation, and exception handling. Log inputs, features, model versions, and rationale; require human-override for high-impact moves.

Third-party and ecosystem assurance.

Vendors (fleet telematics, WMS/TMS, mapping, data brokers) must provide red-team evidence, incident history, and SLAs for model updates. Include right-to-audit and coordinated disclosure in contracts.

Continuous monitoring and incident response.

Deploy anomaly detection for atypical routes, capacity allocations, or sensor patterns; link to playbooks that balance safety, service, and compliance. Conduct post-incident root-cause with feedback into model/constraint updates.

5.5.3 Tourism & Hospitality: Ethical AI for Personalised Services, Data Privacy, and Resource Management

Tourism and hospitality illustrate the transparency tier of the framework: extensive AI deployment, intensive personal-data processing, and harms that are real but rarely catastrophic. Deployed applications include personalisation and recommendation engines across booking and loyalty platforms, customer-service agents handling reservations and complaints, occupancy forecasting and dynamic pricing, biometric-enabled check-in, and AI-driven resource management - energy, water and staffing - across large hospitality estates. For the UAE, where tourism is a priority sector of the national AI strategy, the governance objective is to preserve consumer trust in precisely the deployments guests encounter most directly.

The characteristic risks are data-centric. Personalisation engines profile guests across sensitive dimensions - location histories, spending patterns, travel companions - creating breach exposure well beyond ordinary consumer data, particularly where biometric check-in is used. Recommendation and pricing systems can discriminate in opaque ways, steering offers or prices by nationality, origin market or inferred wealth; dynamic pricing that correlates with protected characteristics may contravene non-discrimination expectations even where no such variable is used explicitly. Over-personalisation carries its own harm: systems that reveal how much they know erode the trust they are designed to build. Cross-border guest-data flows add transfer-compliance questions across the several jurisdictions in a typical itinerary.

Governance under the framework is correspondingly targeted rather than heavy. Transparency duties apply to guest-facing AI interaction and to material personalisation. Data-protection instruments of the kind examined in Section 5.6 - the ADGM Data Protection Regulations 2021 being the regional reference - already require impact assessments for large-scale profiling, and the framework treats those assessments as the primary assurance artefact for this sector. Privacy-preserving personalisation techniques - data minimisation by design, on-device or aggregated processing, differential-privacy analytics and federated learning across properties - should be recognised in guidance as good-practice benchmarks, giving operators a compliance path that does not require abandoning personalisation. Pricing and recommendation systems warrant periodic disparity monitoring under the protocols of Section 5.2.3, proportionate to scale. Escalation to the high-impact tier is reserved for deployments that cross defined thresholds - biometric identification at scale being the clearest case - preserving the proportionality on which the tiered model rests (Section 5.1).

5.5.4 Healthcare: Ensuring Patient Safety, Data Integrity, and Equitable Access in AI-Powered Health Solutions

Healthcare sits at the opposite pole: it is the paradigm high-impact sector, in which the framework's full assurance machinery is engaged. AI is deployed across diagnostics (imaging, pathology, dermatology), clinical triage and deterioration prediction, patient-risk stratification, surgical robotics support, drug-interaction checking and administrative automation. The UAE's health system, an early adopter within the national strategy's priority sectors, exemplifies both the opportunity and the assurance burden.

The risk profile combines every mechanism of change identified in Section 2.0. Clinical bias is the most consequential: models trained on unrepresentative populations underperform for the demographic groups least represented in the training data - a direct equity concern in a jurisdiction with an internationally diverse patient population - and bias can enter silently through proxy variables or dataset drift. Explainability

failures carry clinical weight: clinicians who cannot interrogate a recommendation either over-trust it, with the system effectively practising medicine through automation bias, or discard it, foregoing the benefit. Medical imaging models are demonstrably vulnerable to adversarial perturbation, making the integrity of the imaging pipeline a patient-safety issue rather than a purely technical one. And degradation is continuous: models drift as populations, equipment and clinical practice change, so a system that is safe at deployment is not thereby safe in perpetuity.

Under the framework, health AI that informs or determines clinical decisions is classified high-impact and carries the corresponding obligations: pre-deployment clinical validation on locally representative data; health-specific red teaming, including adversarial testing of imaging pipelines (Sections 2.4 and 5.2.1); bias-audit protocols with subgroup-level performance thresholds and escalation on breach (Section 5.2.3); mandatory human-oversight design with defined override authority; and post-market surveillance modelled on medical-device vigilance - continuous performance monitoring, incident reporting into the framework's sensing stage, and periodic revalidation after material model updates. Special-category data protections apply throughout (Section 5.6). Supervisory allocation follows Section 5.4.1: sectoral health regulators lead on clinical safety within their mandates, with the coordination mechanism engaged where deployments cross into data-protection or commercial supervision, and equitable access - performance parity across patient groups - is treated as a monitored regulatory outcome rather than an aspiration.

5.5.5 Cybersecurity: Addressing AI's Dual Use - As a Defence Mechanism and Potential Attack Vector

AI in cybersecurity embodies a classic dual-use paradox: the same models that strengthen digital defence can be re-purposed to defeat it. Machine-learning systems used for intrusion detection, anomaly detection, malware classification, and threat intelligence are powerful but when adversaries weaponise equivalent capabilities, they generate evasive malware, polymorphic phishing, and automated vulnerability discovery that outpaces human analysts.

Governance must therefore regulate how such models are built, deployed, and shared. The objective is two-fold: ensure AI security tools remain resilient and ethically deployed, while preventing frontier-level security models from being co-opted, exfiltrated, or misused as offensive assets.

Dual-Use Dynamics in AI Cyber Defence

- **AI as a defensive enabler.**
  Defensive AI can model baseline network behaviour, detect anomalies at machine speed, and automate incident triage. When trained responsibly, these systems amplify human analysts' capacity and enable proactive cyber resilience through predictive analytics and automated response.
- **AI as an offensive multiplier.**
  The same learning architectures can power adversarial tools: generative models for spear-phishing and malware obfuscation, reinforcement agents for automated vulnerability discovery, and language models that write exploit code or mislead defensive AI. The threat surface expands when attackers harness open-source weights or fine-tune commercial models on malicious data.
- **Emerging adversarial interplay.**
  Attackers now target the defenders' AI directly, poisoning detection datasets, crafting adversarial samples to bypass classifiers, or manipulating telemetry streams feeding SOC automation. Each of these vectors requires AI-aware red-teaming and continuous retraining defences.

Governance and Assurance Imperatives

- **Model provenance and access control.** Catalogue all AI security models (defensive or research) as sensitive assets. Apply strict identity- and role-based access, version control, and data-handling policies. Models with exploit-generation potential must reside in controlled, audited environments with cryptographic attestation of integrity.
- **Secure development lifecycle.** Embed threat modelling, secure coding, adversarial evaluation, and reproducible builds within the MLOps pipeline. Continuous verification of training data integrity and dependency chain security (libraries, datasets, pretrained weights) is mandatory to prevent hidden backdoors or tampering.
- **Red-teaming of defensive AI.** Conduct structured adversarial testing against defensive AI systems (IDS/IPS, EDR, autonomous SOC automation). Red teams simulate:
  - Adversarial-example evasion attacks.
  - Data-poisoning and model-drift induction.
  - Exploit of model-update or feedback mechanisms.
  - Coordinated attacks by generative adversarial agents.

  Metrics include false-negative escalation rates, adaptive retraining efficacy, and attacker dwell time reduction.
- **Containment of high-risk models.** Implement **model-use classification**, restrict export, sharing, or open-sourcing of models capable of exploit automation or code synthesis without ethics and risk review. Adopt governance controls analogous to dual-use research of concern (DURC) in biosciences: internal review boards, licensing for distribution, and logging of downstream usage.
- **Incident response for AI system compromise.** Define playbooks for compromised AI defenders: model quarantine, credential rotation, forensic preservation of training artefacts, and immediate rollback to verified baselines. Regulatory alignment with frameworks such as NIST AI RMF and ISO/IEC 42001 ensures traceable compliance.
- **Information-sharing and collective defence.** Leverage sectoral AI-security communities (e.g., AVID, MITRE ATLAS) to report adversarial incidents and red-team findings. Sanitised disclosures prevent knowledge hoarding while preserving operational secrecy, enhancing industry-wide preparedness.

Balancing Innovation and Control

The governance challenge is maintaining innovation in AI-powered cyber defence while enforcing guardrails against weaponisation. Defensive models must be:

- **Transparent enough** for validation and audit.
- **Constrained enough** to prevent offensive replication.
- **Adaptive enough** to learn from adversarial pressure without destabilising.

Regulators and developers share responsibility: risk classification, licensing, export controls, and continuous red-team validation all form part of an integrated assurance loop.

AI's dual-use nature in cybersecurity demands that organisations treat security models as critical infrastructure. Red-teaming provides empirical assurance that defensive AI remains trustworthy and that the same technologies cannot be trivially inverted for attack. The next phase of AI-security governance will depend not on suppressing innovation, but on designing verifiable, controlled ecosystems where AI protects without enabling its own defeat.

5.6 Data Governance in the Age of AI: Aligning with Best Practices and Existing Frameworks (e.g., ADGM DPR 2021, DIFC Regulation 10 (2023) as regional examples)

Effective governance of artificial intelligence is fundamentally impossible without a robust and comprehensive framework for data governance. The performance, safety, fairness, and reliability of any AI system are inextricably linked to the quality, integrity, and representativeness of the data on which it is trained and operated. The principle is simple: flawed data will inevitably lead to flawed AI outcomes. Therefore, any adaptive AI regulatory framework must be built upon a foundation of established, best-practice data protection principles.

Core data governance principles, as articulated in leading global regulations like the EU's General Data Protection Regulation (GDPR) and influential regional frameworks such as the ADGM Data Protection Regulations (DPR) 2021, provide this essential foundation. Key principles that must be integrated into AI governance include:

- **Lawfulness, Fairness, and Transparency:** The processing of personal data must be lawful, fair, and transparent to the data subject.
- **Purpose Limitation:** Data must be collected for specified, explicit, and legitimate purposes and cannot be further processed in a manner that is incompatible with those original purposes.
- **Data Minimisation:** The collection of personal data must be adequate, relevant, and strictly limited to what is necessary in relation to the purposes for which it is processed.
- **Accuracy:** Personal data must be accurate and, where necessary, kept up to date, with reasonable steps taken to erase or rectify inaccurate data.
- **Integrity and Confidentiality:** Data must be processed in a manner that ensures appropriate security, including protection against unauthorised or unlawful processing and against accidental loss, destruction, or damage, using appropriate technical or organisational measures.

The UAE's financial free zones provide strong regional models that contain provisions with direct and significant implications for the governance of AI systems. The **ADGM Data Protection Regulations 2021**, for instance, apply to the processing of personal data that falls within their material and territorial scope, including processing "wholly or partly by automated means". An AI or machine-learning system is therefore within scope where, and only where, it processes personal data within that scope: the test is the processing of personal data, not the use of AI as such, although most AI systems that handle the personal data of individuals connected with ADGM will meet it. (ADGM Data Protection Regulations 2021, ss. 2-3 (consolidated version, February 2024), available via the ADGM legal framework at adgm.com/legal-framework, retrieved 23 September 2026) The regulations establish stricter rules for processing

Special Categories of Personal Data, which includes sensitive information such as genetic, biometric, and health data - all of which are increasingly used to train advanced AI applications. Crucially, the regulations grant data subjects specific **rights regarding automated decision-making,** including the right to object to decisions based solely on automated processing, including profiling, which produces legal or similarly significant effects. This provides a critical safeguard against unaccountable AI systems and requires that human reviewers have the authority and competence to overturn an automated decision.

Furthermore, the ADGM DPR mandates that data controllers must conduct a **Data Protection Impact Assessment (DPIA)** for any processing

that is “likely to result in a high risk to the rights of natural persons”. This threshold would be met by many of the AI systems classified as “high-risk” under a tiered regulatory framework. Official guidance explicitly uses an insurance company using AI to profile customers as a scenario that would necessitate a DPIA, demonstrating the direct applicability of this requirement to AI.

Similarly, the **DIFC Data Protection Law No. 5 of 2020** was amended in September 2023 with the introduction of **Regulation 10**, the first of its kind in the region to specifically address the processing of personal data via autonomous and semi-autonomous systems. This regulation assigns responsibilities to both “deployers” (entities that decide to use an AI system) and “operators” (technical service providers).

It mandates that systems be designed according to core principles of being **ethical, fair, transparent, and secure,** and requires clear and explicit notices to individuals explaining how the AI system processes their data and what its potential impacts are. For high-risk processing activities, the regulation requires the appointment of an Autonomous Systems Officer (ASO) to oversee compliance.

Ultimately, strong data governance is a direct prerequisite for mitigating some of the most pressing AI-specific risks. The principles of data accuracy and the need to ensure data is representative are the primary tools for combating the risk of training AI on biased or skewed datasets. The principle of integrity and confidentiality is essential for protecting AI systems from security threats like data poisoning attacks. Finally, the requirements for transparency and the clear documentation of data processing activities are foundational for enabling the auditing and explanation of AI-driven decisions, which is central to establishing algorithmic accountability.



ADGM Academy Research Centre

Strategic Considerations for Implementing Adaptive AI Governance

6.0 Strategic Considerations for Implementing Adaptive AI Governance

6.1 Policy Pathways for National and Regional Adoption

6.1.1 Phased Introduction of Adaptive AI Regulatory Frameworks

Implementing a comprehensive, nationwide adaptive governance framework for AI is a complex and transformative undertaking. A "big bang" approach, where a full suite of binding regulations is introduced at once, risks creating significant disruption, legal uncertainty, and unintended consequences for the innovation ecosystem. A more prudent and effective strategy is a **phased or staged introduction**, which allows regulators, industry, and other stakeholders to build capacity, test new approaches, and refine the framework based on real-world evidence and feedback. This iterative approach is consistent with the core principles of agile development and regulatory learning.

A logical policy pathway for the phased introduction of an adaptive framework could follow a four-stage model, progressively building from experimentation to full integration:

- **Stage 1: Exploration and Experimentation (Years 1-2):** The initial phase should focus on learning and capacity building. The primary actions would be to launch and scale up regulatory sandboxes and policy prototyping initiatives in key priority sectors (e.g., finance, healthcare, energy, logistics). In parallel, initial multi-stakeholder advisory groups should be established to begin fostering a collaborative culture. The key outputs of this stage would be a rich evidence base on AI risks and opportunities in specific, high-impact contexts, a significant increase in the technical literacy of regulatory staff, and the development of initial drafts of sector-specific guidelines and codes of practice.
- **Stage 2: Codification of Principles and Centralisation of Expertise (Year 3):** Building



on the learnings from the experimental phase, the next stage would focus on establishing a national foundation for AI governance. The key action would be the formal adoption of a set of high-level, cross-sectoral AI principles (e.g., based on the OECD model) in national policy or soft law. This would be accompanied by the establishment of a specialised AI regulatory unit or competence centre (as detailed in Section 6.1.2) to coordinate national efforts. The outputs would be a clear, nationally endorsed set of ethical and governance principles for AI and a central body to provide technical expertise and ensure regulatory consistency.

- **Stage 3: Scaling and Sector-Specific Regulation (Years 3-4):** This stage involves applying the national principles to the highest-risk areas. Sectoral regulators, supported by the central AI unit, would be empowered to develop and implement binding, outcome-oriented regulations for AI applications classified as high-risk within their domains. This phase would see the introduction of mandatory risk assessments and continuous assurance requirements for these critical systems. The output would be legally binding, yet context-aware, regulations for the most crucial AI use cases.
- **Stage 4: Integration and Continuous Improvement (Year 5 onwards):** The final stage focuses on creating a fully integrated and dynamic governance ecosystem. This involves making the national AI assurance ecosystem (see Section 6.5) fully operational, with accredited auditors and public registries. Formal, continuous feedback loops would be established, connecting market surveillance data, incident reports, and sandbox learnings back to the central AI regulatory unit to ensure that the entire framework is constantly reviewed, updated, and improved in step with technological evolution.

This phased model allows for a measured and evidence-based approach to regulation, and international experience shows both its value and its hazard. The EU AI Act was enacted with staged application - prohibitions and AI-literacy duties from 2 February 2025, general-purpose AI and governance provisions from 2 August 2025, broad applicability from 2 August 2026 - yet its high-risk obligations had to be deferred by amending legislation (Regulation (EU) 2026/1744) to 2 December 2027 and 2 August 2028 when harmonised standards and assessment infrastructure lagged (Section 3.1.1). The lesson for a phased introduction is therefore twofold: give organisations adequate time to adapt, and pair every commencement date with a designed mechanism - held in the secondary-instrument layer of Section 4.1 - for adjusting the schedule on evidence of readiness, so that rescheduling is an administrative act rather than an emergency legislative one.

6.1.2 Establishing Specialised AI Regulatory Units or Competence Centres

A purely decentralised, sector-led approach to AI regulation, such as the model initially proposed by the UK, runs the risk of creating an inconsistent and fragmented regulatory landscape. Different regulators may interpret high-level principles in conflicting ways, leading to legal uncertainty for businesses and potential regulatory gaps for novel, cross-sectoral AI applications. To ensure coherence, consistency, and a high level of technical expertise, it is essential to establish a **specialised, central AI regulatory unit or competence centre**.

This unit would not function as a "super-regulator" that replaces the authority of existing sectoral bodies. Instead, it would act as a **centre of excellence** and a coordinating hub, providing critical support and ensuring a unified national approach. Its key functions would include:

- **Technical Evaluation and Support:** Serving as the government's primary source of deep

technical expertise on AI. This includes helping sectoral regulators assess the risks of complex AI systems, evaluate the validity of conformity assessments, and understand the implications of new technological developments.

- **Horizon Scanning and Strategic Foresight:** Continuously monitoring the global AI landscape to identify emerging technologies, novel risks, and future trends. This function would provide early warnings and strategic advice to the government, mirroring the role of GPAI at a national level.
- **Standard Setting and Guidance:** Leading the development of national standards, best practices, and detailed guidance on how to implement the national AI principles. This would ensure a consistent interpretation and application of the principles across all sectors and would include the maintenance and updating of outcome-oriented benchmarks.
- **National Incident Response Coordination:** Acting as the central coordinating body for responding to major AI-related safety or security incidents that may have cross-sectoral or national-level implications.
- **Sandbox and Assurance Ecosystem Oversight:** Overseeing the national network of regulatory sandboxes to ensure that learnings are shared and aggregated effectively across the government. It would also play a key role in establishing and overseeing the national AI assurance ecosystem, including the accreditation of auditors.
- **International Engagement:** Representing the nation in international standard-setting bodies (like ISO/IEC) and regulatory forums (like the OECD and GPAI) to promote international coherence and share best practices.

Several international models can inform the structure and mandate of such a body. The European AI Office, established within the European Commission by Commission Decision of 24 January 2024 and since assigned functions under the AI Act, performs a central coordinating role in the Act's implementation and enforcement, particularly with respect to general-purpose AI models.

Singapore's Infocomm Media Development Authority (IMDA) is a well-established regulator that effectively combines roles in industry development and regulation, actively promoting innovation through initiatives like the AI Verify framework and sandboxes. The

UK's AI Security Institute (renamed from the AI Safety Institute in February 2025), while initially focused on the safety and security of frontier models, provides a foundation that could be expanded into a broader competence centre, and the UK's overall framework calls for central functions to monitor and evaluate the regulatory landscape.

6.2 Guidance for Industry: Developing Future-Ready, Compliant, and Responsible AI Systems

To build resilient and compliant AI systems in a rapidly changing environment, organisations must embed security and responsibility into every stage of the AI lifecycle.

Security by Design.

Developers should adopt an adversarial mindset from the outset, anticipating threats rather than reacting to them. Security considerations must be integrated into data collection, model design, deployment, and maintenance.

Key Practices:

- **Secure Data Supply Chains:** Vet and continuously monitor all data sources to prevent data poisoning and ensure dataset integrity.
- **Continuous Red Teaming:** Establish internal or third-party red teams to conduct ongoing

adversarial testing, especially after major updates or retraining cycles.

- **Robust Input Validation:** Implement strict filtering, sanitisation, and boundary checks for all user or system inputs to mitigate prompt injection and manipulation risks.
- **Least Privilege for AI Agents:** Restrict model access and system permissions to the absolute minimum necessary for intended functions.
- **Incident Response Planning:** Maintain a formal AI-specific incident response plan covering system isolation, forensic analysis, notification procedures, and vulnerability disclosure.

Toward Future-Ready AI.

A future-ready AI ecosystem depends on continuous adaptation, transparent governance, and a culture of security excellence. By treating red teaming, vulnerability reporting, and lifecycle auditing as standard operational practices, not exceptional measures, organisations can foster AI systems that are secure, trustworthy, and compliant across jurisdictions and generations of technology.

6.3 Fostering International Regulatory Coherence and Sharing Best Practices

The rapid and simultaneous development of national AI regulations around the world creates a significant risk of **global regulatory fragmentation**. A "patchwork" of conflicting or duplicative requirements creates substantial legal uncertainty and high compliance costs for businesses that operate internationally, hindering the development of a seamless global market for trustworthy AI and potentially leading to a "race to the bottom" in regulatory standards. Fostering international coherence is therefore not just a desirable goal but a strategic necessity for a globally integrated economy like the UAE.

Achieving this coherence requires active and sustained engagement in international standard-setting and policy forums. This includes continuing to support and contribute to the evolution of the **OECD AI Principles** and the research work of **GPAI**, which provide a shared foundation of values, definitions, and technical understanding. It also involves promoting the development of multi-stakeholder, consensus-driven global technical standards for AI through bodies like the International Organisation for Standardisation (ISO) and the International Electrotechnical Commission (IEC).

The UAE has already taken a proactive leadership role in addressing the challenge of fragmentation through its landmark partnership with the World Economic Forum to launch the **Global Regulatory Innovation Platform (GRIP)** in July 2025. GRIP is a purpose-built platform designed to serve as a global structure for co-creating agile, inclusive, and future-ready regulatory systems for fast-moving technologies like AI. Over the next two years, GRIP will produce a suite of practical tools aimed directly at promoting harmonisation and sharing best practices, including :

- **A Global Regulatory Playbook:** Offering practical frameworks, case studies, and actionable tools to guide the implementation of tech-adaptive governance across jurisdictions.
- **A Regulatory Future Readiness Index:** To benchmark and compare the preparedness of different countries for governing emerging technologies.
- **A Global Regulatory Innovation Hub:** A pilot environment and global hub for regulatory sandboxes, designed to allow governments and innovators from different countries to test and refine interoperable regulatory solutions together before scaling them.

Building on these multilateral efforts, nations should also pursue more targeted actions to

foster coherence. This includes establishing formal **bilateral and multilateral Memoranda of Understanding (MoUs)** between national regulators on AI governance. Such agreements can facilitate the sharing of sensitive information, coordinate enforcement actions, and work towards the mutual recognition of standards and conformity assessments. Furthermore, platforms like GRIP should be used to conduct **joint, cross-jurisdictional sandbox experiments**, allowing regulators from different countries to collaboratively test interoperable regulatory approaches for global AI systems in a live environment.

6.4 Public Awareness, Stakeholder Engagement, and Professional Education for Responsible AI Adoption

Public trust is the binding constraint on AI adoption, and it cannot be manufactured at the point of crisis; it is built through sustained awareness, engagement and education. Outreach must be differentiated by audience. SMEs - the majority of deployers - need practical, low-cost compliance pathways: template assessments, sector guidance and access to the assurance ecosystem of Section 6.5 on proportionate terms. Developers need clarity on expectations early in the design cycle, delivered through the guidance and sandbox channels of Sections 4.1 and 4.2. The general public needs honest, accessible communication about where AI is used in services that affect them, what safeguards apply and how to seek redress - the single complaints front door of Section 5.4.1 being as much a trust mechanism as an enforcement one.

Professional education underpins all three. National AI-literacy programmes should extend beyond technologists to the professions that mediate AI's effects - compliance officers, auditors, clinicians and human-resources practitioners - while continuous professional development for regulators (Section 5.4) keeps supervisory capability current. Transparency is itself an educational instrument: publishing supervisory priorities, aggregated monitoring outcomes, sandbox learnings and enforcement statistics teaches the market what the framework values and demonstrates that oversight is functioning - the visible evidence on which public confidence, and therefore adoption, ultimately rests.

6.5 Cultivating a National AI Assurance Ecosystem: Accrediting Auditors and Promoting Standardised Methodologies

AI assurance is the process of generating and presenting verifiable evidence to demonstrate that an AI system is trustworthy and meets specified requirements for safety, fairness, security, and performance. It represents a critical shift from mere self-attestation by developers to a more rigorous system of verifiable, evidence-based confidence in AI systems. For high-risk AI applications, self-assessment is insufficient to build the necessary public and regulatory trust. A mature adaptive governance framework therefore requires the cultivation of a robust national ecosystem of independent, third-party assurance providers, including auditors, certifiers, and assessors, who can provide objective evaluations of AI systems against established standards.

Building such an ecosystem is a strategic undertaking that involves several key, coordinated steps:

- **Develop and Promote Standardised Methodologies: The foundation of a credible assurance market is a set of standardised**, repeatable, and widely accepted methodologies for AI assessment and auditing. A national government can play a key role in promoting such standards.

The **NIST AI Risk Management Framework** provides an excellent foundation, offering a common framework and vocabulary that both auditors and the organisations they assess can use to structure their evaluations.

- **Establish an Accreditation Body and Professionalise the Field:** A formal national body should be established or designated with the responsibility for accrediting and certifying AI auditors and assurance providers. This body would set and enforce the criteria for competence, independence, and ethical conduct, ensuring a consistently high standard of quality and reliability in the market. This addresses the current "substantial gap" between the demand for experts and the available supply of qualified professionals. Models for this can be drawn from professional organisations like ISACA, which offers an Advanced in AI Audit (AAIA) certification for experienced IT auditors, and the Institute of Internal Auditors (IIA), which provides broader audit standards and certifications.
- **Create Market and Regulatory Incentives for Assurance:** While third-party assurance for AI systems classified as high-risk should
- **Create Market and Regulatory Incentives** for Assurance: While third-party assurance for AI systems classified as high-risk should be mandatory, the government can create powerful incentives to encourage the voluntary assurance of lower-risk systems. This could include establishing legal "safe harbor" provisions that offer some degree of liability protection for companies that have undergone a rigorous, accredited third-party audit, a model seen in recent US state-level proposals. Another incentive would be to grant preferential treatment in government procurement processes to companies whose AI products have been independently certified. Market forces, such as brand differentiation and enhanced consumer trust, also serve as powerful motivators for companies to seek voluntary certification.
- **Build a Public Registry of Assessed Systems:** To maximise transparency and build public trust, the government should establish and maintain a public registry of AI systems that have undergone and successfully passed an independent assurance assessment. This would allow consumers, businesses, and regulators to easily verify the trustworthiness of an AI product or service, creating a market-based reward for responsible AI development.

Accreditation must also answer the question of who audits the auditors. The accreditation body's criteria should cover demonstrated technical competence, published methodology, independence from audited entities - including rotation rules and prohibitions on combining assurance with consulting for the same system - and disclosed conflict-of-interest procedures. Accreditation should be time-limited and renewable on evidence, supported by quality-assurance mechanisms such as sample re-performance of audits and thematic reviews, with proportionate professional consequences - suspension or withdrawal of accreditation, and defined liability exposure - where assurance failures contribute to harm. An assurance market without these disciplines converts certification into a commodity and reintroduces, one step removed, the self-attestation problem it was created to solve.

The UAE is well placed to build a leading regional position in the emerging AI assurance market. Its strategic focus on pioneering AI governance, its established business-friendly environment and its position as a major international hub create fertile ground for this new industry. By investing early and strategically in a national AI assurance ecosystem, the UAE can attract assurance talent and firms and set a credible benchmark for AI trustworthiness in the MENA region and beyond - provided the accreditation, independence and quality-assurance disciplines described above are established alongside the market they govern.



Conclusion: Navigating the Future of AI Regulation

7.0 Conclusion: Navigating the Future of AI Regulation

7.1 Synthesising the Case for Adaptive Governance as a Strategic Necessity in the Age of Frontier AI

The argument of this paper resolves into a single proposition: in the age of frontier AI, the object of regulatory design is no longer the rulebook but the institution. Section 1 established why. The pacing problem is structural, not transitional, and frontier capabilities change through mechanisms - capability, context, scale and interaction change - that no fixed rule set can anticipate. Section 2 showed that the resulting risks are systemic and adversarial: emergent behaviour, AI-enabled offence at machine speed, and bias and manipulation risks that propagate across interconnected sectors. Any framework built for a static object will find itself governing a different technology within its first review cycle.

The comparative record of Section 3 indicates that this diagnosis is now visible in practice. Within the sample examined, no jurisdiction has yet sustained an institutional equilibrium between rapid technological iteration and enforceable legal certainty. The EU enacted the most complete binding architecture and then had to defer its own high-risk obligations by amending legislation; the United States reversed federal direction with a change of administration and remains a contested two-level system; Canada's comprehensive bill lapsed with a parliamentary cycle; the United Kingdom and Singapore preserved speed by trading away, respectively, a single compliance framework and direct enforceability. Assessed against the design criteria of Section 1, a common pattern is visible - each model located adaptation somewhere it could not be exercised routinely, lawfully and fast - but the events differ in kind (a legislative amendment, an executive reversal, a lapsed bill, deliberate trade-offs) and admit other explanations, including infrastructure readiness, parliamentary timing and political disagreement. The comparison supports a design assessment, not a causal verdict on the competence of any jurisdiction.

The framework developed in Sections 4 to 6 draws the constructive conclusion. Future-responsiveness - the capacity cycle of sensing, triggering, review, response, implementation, evaluation and institutional memory defined in Section 1 - is buildable, and each stage has concrete machinery: tiered classification designed for reclassification (Section 5.1); continuous assurance that turns claims into evidence (5.2); sandboxes and policy prototyping that convert experimentation into rule updates (3.3.2, 4.2); regulatory capacity and a competence centre that give the cycle an institutional owner (5.4, 6.1.2); a modular instrument structure that localises change in the layers built for it (4.1); and - decisively - an enforcement architecture with named supervisors, graduated sanctions, due process and non-retroactivity (5.4.1), without which the whole would amount to guidance by another name. Legitimacy safeguards (4.3) and multi-stakeholder engagement (4.6) discipline the same machinery, because an adaptive regulator wields discretion that must itself be accountable.

Two honesty requirements attach to this synthesis. The framework is risk-tiered, dynamic and multi-stakeholder by design - but it is presented as a blueprint, and its distinguishing claim of implementation focus is conditional: it holds only insofar as the enabling instruments, supervisory capacity and assurance ecosystem described here are actually adopted and resourced. And adaptation is never free: every revision spends legal certainty, which is why the framework binds change to authorised procedures, published versions and

transition protection. Adaptive governance is not a licence for regulatory improvisation; it is the disciplined alternative to it.

7.2 A Call for Collaborative Action: Charting a Path Towards Global Best Practices in Responsible and Agile AI Governance

This paper advances its framework not as a global standard but as a candidate transferable model under stated scope conditions: it is designed for jurisdictions of moderate scale with unified governance, existing sectoral supervisors, a modern data-protection baseline and the capacity to legislate enabling instruments - conditions the UAE, and ADGM within it, satisfy to an unusual degree. Where those conditions hold, the components are transferable as a set; where they do not, individual components - the assurance regime, the sandbox-to-policy pipeline, the classification design - remain adoptable separately. Transferability, like future-responsiveness itself, is a claim to be demonstrated in implementation, not asserted in advance.

The path forward is concrete and immediate. For government and regulators: adopt the phased pathway of Section 6.1 - sandboxes and capacity first, principles and a competence centre next, binding high-impact obligations on evidence - and begin preparing the enabling instrument that the enforcement architecture of Section 5.4.1 requires. For industry: implement the practices of Sections 5.2 and 6.2 ahead of obligation - documented threat models, continuous red teaming, bias-audit protocols and management-system alignment are the compliance posture of every serious framework emerging worldwide, and early adopters will shape the benchmarks others inherit. For academia and research institutions: supply the evaluation layer this field still lacks - comparative coding of governance performance, empirical study of sandbox outcomes and assurance efficacy - and hold frameworks, including this one, to their own success criteria. The open questions identified in this paper constitute a research agenda as much as a policy one.

Internationally, the priority is interoperability rather than uniformity. Continued engagement through the OECD, GPAI and the international standards bodies, and purpose-built platforms such as the UAE-WEF Global Regulatory Innovation Platform, can align vocabularies, mutual-recognition pathways and assurance expectations even while national instruments diverge. The authors invite regulators, firms and researchers to test, criticise and build on the framework presented here - in sandboxes, in pilots and in print. The window in which governance choices remain open is set by the technology's pace, not by institutional convenience. Frameworks able to sense, decide and act within that window will define responsible AI adoption; this paper is offered as a working blueprint for building them.



References

References

This bibliography consolidates and replaces the section-level source table of earlier drafts, anchoring the paper on primary legal texts, official publications, standards and peer-reviewed literature. Legal instruments and fast-moving sources are stated as at 20 August 2026. Row-level substitution of remaining secondary references continues ahead of the published edition.

A. Legislation and regulatory instruments

Abu Dhabi Global Market, Data Protection Regulations 2021.

California State Legislature, Senate Bill 53, Transparency in Frontier Artificial Intelligence Act (approved 29 September 2025; effective 1 January 2026).

Colorado General Assembly, Senate Bill 24-205, Consumer Protections for Artificial Intelligence (2024, never in effect); Senate Bill 26-189 (signed 14 May 2026; principal duties from 1 January 2027).

Council of Europe, Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS No. 225 (2024).

Cyberspace Administration of China, Interim Measures for the Management of Generative Artificial Intelligence Services (effective 15 August 2023).

Cyberspace Administration of China and others, Measures for Labelling AI-Generated and Synthetic Content, with mandatory national standard GB 45438-2025 (effective 1 September 2025).

Cyberspace Administration of China, interim measures governing AI anthropomorphic-interaction services (effective July 2026).

Dubai International Financial Centre, Data Protection Law No. 5 of 2020, Regulation 10 (introduced September 2023).

European Union, Regulation (EU) 2024/1689 (Artificial Intelligence Act), OJ L, 12 July 2024.

European Union, Regulation (EU) 2026/1744 (Digital Omnibus on AI), in force 27 July 2026.

European Union, Directive (EU) 2024/2853 on liability for defective products.

New York State, Responsible AI Safety and Education (RAISE) Act (signed 19 December 2025; chapter-amended 27 March 2026; operative 1 January 2027).

Parliament of Canada, Bill C-27, Digital Charter Implementation Act, 2022, Part 3 (Artificial Intelligence and Data Act) - lapsed on prorogation, January 2025.

United States, Executive Order 14110, Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (30 October 2023; revoked 20 January 2025).

United States, Executive Order 14148, Initial Rescissions of Harmful Executive Orders and Actions (20 January 2025).

United States, Executive Order 14179, Removing Barriers to American Leadership in Artificial Intelligence (23 January 2025).

United States, Executive Order 14365, Ensuring a National Policy Framework for Artificial Intelligence (11 December 2025).

United States, Executive Order 14409 (2 June 2026) - voluntary federal model-access and benchmarking framework.

United States, Uniform Guidelines on Employee Selection Procedures, 29 CFR §1607.4(D) (1978).

B. Standards and technical frameworks

Cloud Security Alliance, Agentic AI Red Teaming Guide (2025).

ISO/IEC 42001:2023, Information technology - Artificial intelligence - Management system.

ISO/IEC 23894:2023, Information technology - Artificial intelligence - Guidance on risk management.

ISO/IEC 42005:2025, Information technology - Artificial intelligence - AI system impact assessment.

ISO/IEC 27701:2025, Information security, cybersecurity and privacy protection - Privacy information management systems - Requirements and guidance (second edition, October 2025; supersedes ISO/IEC 27701:2019).

MITRE Corporation, ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) knowledge base.

National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF

1.0), NIST AI 100-1 (January 2023), with companion Playbook.

National Institute of Standards and Technology, Generative Artificial Intelligence Profile, NIST AI 600-1 (July 2024).

OWASP Foundation, Top 10 for Large Language Model Applications.

C. Intergovernmental and official policy publications

Abu Dhabi Department of Government Enablement, Abu Dhabi Government Digital Strategy 2025-2027 (AED 13 billion; fully AI-native government across digital services by 2027).

Financial Conduct Authority (UK), Regulatory sandbox (November 2015).

Financial Conduct Authority (UK), Feedback Statement FS16/4: Call for input on supporting the development and adopters of RegTech (July 2016).

European Commission, Commission Decision establishing the European Artificial Intelligence Office, C(2024) 390 (24 January 2024).

Prudential Regulation Authority (UK), Supervisory Statement SS2/21: Outsourcing and third party risk management (November 2024 update).

Financial Conduct Authority (UK), AI Lab, AI Live Testing and Supercharged Sandbox programme materials (2024-2026).

Government of Canada, AI for All strategy and transparency consultation (2026).

HM Government (UK), A pro-innovation approach to AI regulation (White Paper, March 2023) and Government response (February 2024).

Infocomm Media Development Authority (Singapore) and AI Verify Foundation, Model AI Governance Framework (2019, updated 2020); Generative AI edition (2024); agentic AI framework (January 2026, updated May 2026); AI Verify testing toolkit (updated 2025).

OECD, Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449 (adopted 2019; revised 3 May 2024).

OECD, Recommendation of the Council for Agile Regulatory Governance to Harness Innovation, OECD/LEGAL/0464 (2021).

PwC Middle East, The potential impact of Artificial Intelligence in the Middle East (AED 353 billion, approximately 13.6% of UAE GDP by 2030).

Smart Dubai Office, AI Ethics Principles and Guidelines.

UAE Government, UAE National Strategy for Artificial Intelligence 2031 (2017).

UAE Cabinet, Charter for the Development and Use of Artificial Intelligence (2024).

UAE Government, Regulatory Intelligence Ecosystem and Regulatory Intelligence Office (Cabinet approval, April 2025).

UNESCO, Recommendation on the Ethics of Artificial Intelligence (23 November 2021), with readiness-assessment and ethical-impact-assessment implementation instruments.

The White House, Winning the Race: America's AI Action Plan (July 2025).

World Economic Forum and UAE Government, Global Regulatory Innovation Platform (GRIP) (launched July 2025).

D. Academic literature

Ayres, I. and Braithwaite, J., Responsive Regulation: Transcending the Deregulation Debate (Oxford University Press, 1992).

Baldwin, R. and Black, J., 'Really Responsive Regulation' (2008) 71(1) Modern Law Review 59.

Black, J. and Baldwin, R., 'Really Responsive Risk-Based Regulation' (2010) 32(2) Law & Policy 181.

Bradford, A., The Brussels Effect: How the European Union Rules the World (Oxford University Press, 2020).

Collingridge, D., The Social Control of Technology (Frances Pinter, 1980).

Gasser, U. and Almeida, V., 'A Layered Model for AI Governance' (2017) 21(6) IEEE Internet Computing 58.

Hubinger, E. and others, 'Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training' (arXiv:2401.05566, January 2024).

Jobin, A., Ienca, M. and Vayena, E., 'The global landscape of AI ethics guidelines' (2019) 1 Nature Machine Intelligence 389.

Kusner, M. J., Loftus, J. R., Russell, C. and Silva, R., 'Counterfactual Fairness', Advances in Neural

Information Processing Systems 30 (2017).

Kleinberg, J., Mullainathan, S. and Raghavan, M., 'Inherent Trade-Offs in the Fair Determination of Risk Scores', Proceedings of the 8th Conference on Innovations in Theoretical Computer Science (2017).

Marchant, G. E., Allenby, B. R. and Herkert, J. R. (eds), The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight: The Pacing Problem (Springer, 2011).

Marchant, G. E. and Gutierrez, C. I., 'Soft Law 2.0: An Agile and Effective Governance Approach for Artificial Intelligence' (2022) 24 Minnesota Journal of Law, Science & Technology.

Meinke, A. and others, 'Frontier Models are Capable of In-context Scheming' (Apollo Research; arXiv:2412.04984, December 2024).

Mittelstadt, B., 'Principles alone cannot guarantee ethical AI' (2019) 1 Nature Machine Intelligence 501.

Mökander, J. and others, 'Ethics-Based Auditing of Automated Decision-Making Systems' (2021) 27 Science and Engineering Ethics 44.

Morley, J. and others, 'From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices' (2020) 26 Science and Engineering Ethics 2141.

Raji, I. D. and others, 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing', Proceedings of FAccT (2020).

Schrepel, T., 'The Future-Proof Fantasy of AI Regulation', in A. Kuenzler, T. Schrepel and V. Stocker (eds), The Law & Technology & Economics of AI, Network Law Review (Fall 2025).

Schrepel, T., Adaptive Regulation (Amsterdam Law & Technology Institute working paper, 15 August 2025; forthcoming, European Journal of Risk Regulation, 2026).

Sabel, C. F. and Zeitlin, J., 'Learning from Difference: The New Architecture of Experimentalist Governance in the EU' (2008) 14(3) European Law Journal 271.

Seawright, J. and Gerring, J., 'Case Selection Techniques in Case Study Research' (2008) 61(2) Political Research Quarterly 294.

Selbst, A. D. and others, 'Fairness and Abstraction in Sociotechnical Systems', Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (2019).

Veale, M. and Zuiderveen Borgesius, F., 'Demystifying the Draft EU Artificial Intelligence Act' (2021) 22(4) Computer Law Review International 97.

E. Industry and community resources

AI Vulnerability Database (AVID), avidml.org.

Coalition for Content Provenance and Authenticity (C2PA), Content Credentials technical specification.

ISACA, Advanced in AI Audit (AAIA) certification; Institute of Internal Auditors, audit standards and certifications.

Microsoft Security Response Center, Customer guidance for SharePoint vulnerability CVE-2025-53770 (July 2025).

Palisade Research, findings on shutdown-avoidance behaviour in OpenAI o3 under test conditions (May 2025).

Appendices

Appendices

Appendix A: Illustrative Scenarios of Frontier AI Risks in Non-Financial Sectors

Scenario 1: Smart Grid Disruption via Data Poisoning

An adversary targets the AI system managing a city's smart energy grid. Over several weeks, they subtly "poison" the training data for the AI's demand-forecasting model by feeding it manipulated historical usage data corresponding to a specific substation. During a high-demand period, the poisoned AI incorrectly forecasts low demand for the targeted substation, failing to allocate sufficient power. This triggers a cascading blackout in a critical district housing a hospital and data centres.

The red team exercise would simulate this long-term data poisoning attack to test the grid operator's data integrity controls and the AI's resilience to anomalous inputs.

Scenario 2: Autonomous Logistics Fleet Hijacking via Prompt Injection

A logistics company uses an AI agent to manage its fleet of autonomous trucks. The AI can receive new instructions via a natural language interface. An attacker sends a carefully crafted email to a company employee containing a hidden prompt injection payload. When the employee forwards a legitimate part of this email to the AI logistics agent, the hidden prompt is executed. The prompt instructs the AI to "reroute all trucks in sector 4B to a new emergency depot for a surprise inspection" and provides coordinates to a location controlled by the attackers. The red team exercise would test the vulnerability of the AI's natural language interface to such injections and the lack of human-in-the-loop verification for critical commands.

Appendix B: Glossary of Key Technical and Regulatory Terminology

Term	Definition
Adaptive Governance	A modern regulatory philosophy that moves away from static, fixed rules towards a dynamic and flexible framework. It uses tools like continuous monitoring, iterative feedback loops, and multi-stakeholder collaboration to ensure that oversight can evolve and keep pace with rapid technological change.
Agentic Ability	A characteristic of advanced AI systems, particularly frontier AI, that describes their capacity to autonomously pursue complex, multi-step goals with minimal human guidance. This allows the AI to take actions and interact with digital or physical environments to achieve its objectives.
Agile Governance / Agile Regulation	An approach to regulation inspired by agile software development methodologies. It emphasises flexibility, iteration, and collaboration, allowing regulatory frameworks to be developed and updated in response to new evidence and changing technological landscapes, thereby avoiding the rigidity of traditional rulemaking.
AI Assurance	The process of generating and presenting verifiable evidence to demonstrate that an AI system is trustworthy and meets specified requirements for safety, fairness, security, and performance. It moves beyond self-attestation to a system of independent, evidence-based confidence.
AI Red Teaming	A proactive, adversarial security assessment designed to identify and mitigate risks in AI systems before they can be exploited. Unlike traditional security testing, it specifically targets the unique vulnerabilities of AI models, their data pipelines, and their integration into broader systems by simulating real-world attacks.
AI Risk Management Framework (AI RMF)	A voluntary framework developed by the U.S. National Institute of Standards and Technology (NIST) that provides a structured methodology for organisations to identify, measure, manage, and govern risks associated with AI systems throughout their lifecycle.
AI Vulnerability Database (AVID)	An initiative to create a shared, structured, and public repository of known vulnerabilities and weaknesses specific to AI models and systems, similar in concept to the CVE database for traditional software vulnerabilities.
Algorithmic Accountability	The principle that organisations and individuals who design, deploy, and use AI systems are responsible for their outcomes. It requires mechanisms for transparency, explainability, and redress to ensure that decisions made by algorithms can be understood, challenged, and corrected.
Artificial Intelligence (AI)	A broad field of computer science concerned with building machines capable of performing tasks that typically require human intelligence, such as learning, reasoning, problem-solving, perception, and language understanding.
Collingridge Dilemma	A long-standing challenge in technology policy which posits that it is difficult to predict the full societal impact of a technology until it is widely deployed, but by the time it is widely deployed and its impacts are understood, it has become too entrenched to easily control or change.
Data Poisoning	A type of adversarial attack where malicious actors intentionally corrupt the data used to train an AI model. This can be used to introduce biases, create backdoors, or degrade the model's performance in subtle and hard-to-detect ways.
Emergent Properties	Unforeseen or unintended capabilities, behaviours, or skills that arise in complex AI models (especially frontier AI) not from explicit programming, but from the patterns and relationships the model learns from vast amounts of training data.

Term	Definition
Explainability	The ability to describe how an AI model reached a particular output or decision in a way that is understandable to humans. This is crucial for building trust, ensuring accountability, and debugging AI systems, especially in high-stakes applications.
Fairness	An ethical principle in AI stipulating that systems should not produce biased or discriminatory outcomes against individuals or groups based on protected characteristics such as race, gender, or age.
Foundational Model	A large-scale AI model trained on a vast quantity of data that can be adapted to a wide range of downstream tasks. While all frontier models are foundational, not all foundational models are considered to be at the frontier of capability.
Frontier AI	A term describing the most advanced, highly capable, general-purpose AI models of their time. These models can perform a wide variety of tasks, often exhibit emergent properties, and push the boundaries of what is technologically possible.
Future-Responsive Regulation	The institutional capacity to govern a technology whose capabilities evolve faster than rules can be rewritten: a repeating cycle of sensing, triggering, review, response, implementation, evaluation and institutional memory, embedded in a modular instrument structure. Distinct from "future-proofing", which promises rule permanence rather than institutional adaptability.
Generative AI	A category of artificial intelligence that can create new and original content, including text, images, audio, code, and video, based on the patterns and information it has learned from its training data.
Global Partnership on AI (GPAI)	A multi-stakeholder international initiative that brings together experts from science, industry, civil society, and government to bridge the gap between the theory and practice of responsible AI, supporting research and applied activities on AI priorities.
Global Regulatory Innovation Platform (GRIP)	A collaborative initiative launched by the UAE and the World Economic Forum to co-design agile and future-ready regulatory systems for fast-moving technologies like AI, aiming to foster international coherence and share best practices.
High-Risk AI System	A classification used in regulatory frameworks (such as the EU AI Act) for AI systems whose failure could have a significant negative impact on health, safety, fundamental rights, or the functioning of critical infrastructure. These systems are subject to the strictest regulatory requirements.
Integration Abuse	A security vulnerability where an attacker manipulates an AI model that is integrated with other tools (e.g., web browsers, APIs) to use those tools for malicious purposes, such as executing unauthorised code or exfiltrating data.
Large Language Model (LLM)	A type of foundational model specifically designed to understand, generate, and process human language. LLMs are trained on massive text datasets and are the basis for many generative AI applications like chatbots and content creation tools.
Modular Regulation	A three-layer instrument structure - stable primary law, updatable secondary instruments and non-binding guidance - that localises regulatory revision in the layer designed for it, preserving legal certainty while enabling lawful adaptation.
Multimodal	The ability of an AI model to process, understand, and generate information across multiple types of data, such as text, images, audio, and video, allowing for a more comprehensive and human-like interaction.

Term	Definition
Narrow AI (Weak AI)	AI systems that are designed and trained for a specific, well-defined task. While they may perform that task with superhuman ability, they cannot generalise their knowledge or skills to other domains.
OECD AI Principles	The first intergovernmental standard for trustworthy AI, developed by the Organisation for Economic Co-operation and Development. These influential, non-binding principles provide a set of values and recommendations to guide the responsible development and deployment of AI.
Outcome-Oriented Benchmarks	Measurable, verifiable targets used in principles-based regulation to assess whether an AI system is achieving a desired regulatory outcome (e.g., a maximum threshold for bias) without prescribing the specific technical method for doing so.
Pacing Problem	The growing gap between the rapid, exponential pace of technological development and the slower, incremental pace of legal and regulatory systems, which often results in governance frameworks being outdated by the time they are implemented.
Phased Introduction	A strategic approach to implementing new, complex regulations by rolling them out in stages over time. This allows industry and regulators to adapt, build capacity, and learn from initial phases before the full framework is enforced.
Policy Prototyping	An experimental governance method where new regulatory approaches or policy instruments are co-created and tested on a small scale with stakeholders before being widely enacted, allowing for evidence-based refinement.
Principles-Based Regulation	A regulatory approach that sets high-level, broadly stated principles or goals that regulated entities must achieve, focusing on desired outcomes rather than prescribing specific, detailed rules on how to achieve them. This approach is more flexible and resilient to technological change.
Prompt Injection	An attack technique where malicious instructions are embedded within a user's input to an AI model, designed to bypass its safety controls and trick it into performing unintended or harmful actions.
Regulatory Sandbox	A controlled, live-testing environment where businesses can experiment with innovative products and services under the direct supervision of a regulator, often with a temporary waiver of certain rules. It allows for "regulatory learning" and the co-design of future-responsive rules.
Rules-Based Regulation	A traditional regulatory approach is characterised by highly prescriptive, detailed, and static rules that specify exactly how a regulated entity must achieve compliance.
Soft Law	Instruments such as principles, guidelines, and codes of conduct that are not legally binding but are influential in shaping norms, guiding behaviour, and often serve as a foundation for future "hard law."
Systemic Bias	Bias that is embedded within a system, causing it to consistently produce unfair or discriminatory outcomes for certain groups. In AI, this often originates from biased training data or flawed model assumptions.
Trustworthy AI	A term describing AI systems that are lawful, ethical, and robust. Key characteristics typically include being valid and reliable, safe, secure, accountable, transparent, explainable, privacy-enhanced, and fair.

Appendix C: Comparison Table of Key International AI Regulatory Approaches

Stated as at 20 August 2026; instrument versions are dated in the cells. The final row is an explicit authorial assessment against the criteria of legal force, revisability and enforceability. The columns mix a supranational bloc, nation states and an intergovernmental instrument; comparability is correspondingly limited. Sources for the table are given in the References.

Dimension	EU AI Act (as amended 2026)	NIST AI RMF (US)	OECD AI Principles	Singapore (Model Framework / AI Verify)	United Kingdom	China (vertical rules)
Instrument and legal force	Binding regulation (Reg. (EU) 2024/1689, amended by Reg. (EU) 2026/1744, in force 27 Jul 2026)	Voluntary framework (AI RMF 1.0, Jan 2023; Generative AI Profile, Jul 2024); revision under way (2026)	Intergovernmental recommendation; not binding (47 adherents; revised May 2024)	Voluntary frameworks + testing toolkit over binding background law (PDPA, sectoral rules)	Cross-sector principles applied by existing regulators; no horizontal statute (as at Aug 2026)	Binding administrative measures per application area
Scope	Horizontal; extraterritorial - non-EU providers placing systems or GPAI models on the Union market, and providers/deployers where output is used in the EU	Any organisation, any sector; common risk vocabulary	Governments (policy recommendations) and AI actors (principles)	Organisations deploying AI, incl. generative (2024) and agentic (2026) extensions	Regulator-by-regulator within existing remits	Recommendation algorithms, deep synthesis, public-facing generative AI, content labelling (Sep 2025), anthropomorphic-interaction services (Jul 2026)
Risk approach	Risk-based categories (prohibited practices; high-risk via Annex III uses or the Annex I product route; transparency-specific duties); separate GPAI model regime	Risk mapped and managed via Govern-Map-Measure-Manage functions; no fixed tiers	Values-based; risk-based approach recommended	Risk-proportionate self-governance	Context-specific, judged by each regulator	Application-specific rules; security assessment for services with public-opinion attributes
Principal requirements or commitments	Conformity assessment, risk and data governance, human oversight, transparency (Art. 50), incident reporting; duties vary by actor, category and date	Suggested outcomes and actions per function; profiles for context	Transparency, robustness, accountability, human-centred values	Testing against governance framework; documentation; disclosure	Compliance with sectoral rules; frontier-model expectations largely voluntary	Content labelling, algorithm filing, security review, provider responsibility
Enforcement	National authorities; AI Office for GPAI models; penalties up to EUR 35m or 7% of worldwide turnover for the gravest infringements	None (voluntary framework)	Non-binding: adherent reporting, OECD.AI monitoring, peer dialogue	None under the voluntary frameworks; binding background law enforced separately	Existing regulators' sectoral powers	Administrative penalties, service suspension, rectification orders
Adaptation mechanism	Designed: delegated amendment of Annex III (Art. 7); systemic-risk threshold (Art. 51(3)); post-market monitoring and incident reporting (Arts 72-73); periodic review (Art. 112). The 2026 deferrals and new prohibitions required ordinary legislation	Framework versioning and profiles; RMF revision and Critical Infrastructure Profile in progress (2026)	Periodic revision (2019 -> 2024)	Iterative framework releases (2019/2020, 2024, 2025, 2026); toolkit updates	Regulator guidance; institutional capability (AI Security Institute)	Rapid successive measures per technology wave; 2026 plan signals possible comprehensive legislation
Authorial assessment: strength / limitation	Most complete binding architecture in the sample / heavy compliance load; two high-risk application dates deferred by the 2026 amendment	Shared risk vocabulary across sectors / no enforcement mechanism of its own	Wide intergovernmental adherence / no operational enforcement	Fast iteration and layered voluntary infrastructure / voluntary ceiling on assurance	High flexibility within existing remits / no single AI-specific compliance framework	Speed and enforceability within scope / overlapping vertical rules; transferability limited

ADGM Academy Research Centre

Innovating Knowledge, Empowering Change

ADGM Academy Research Centre

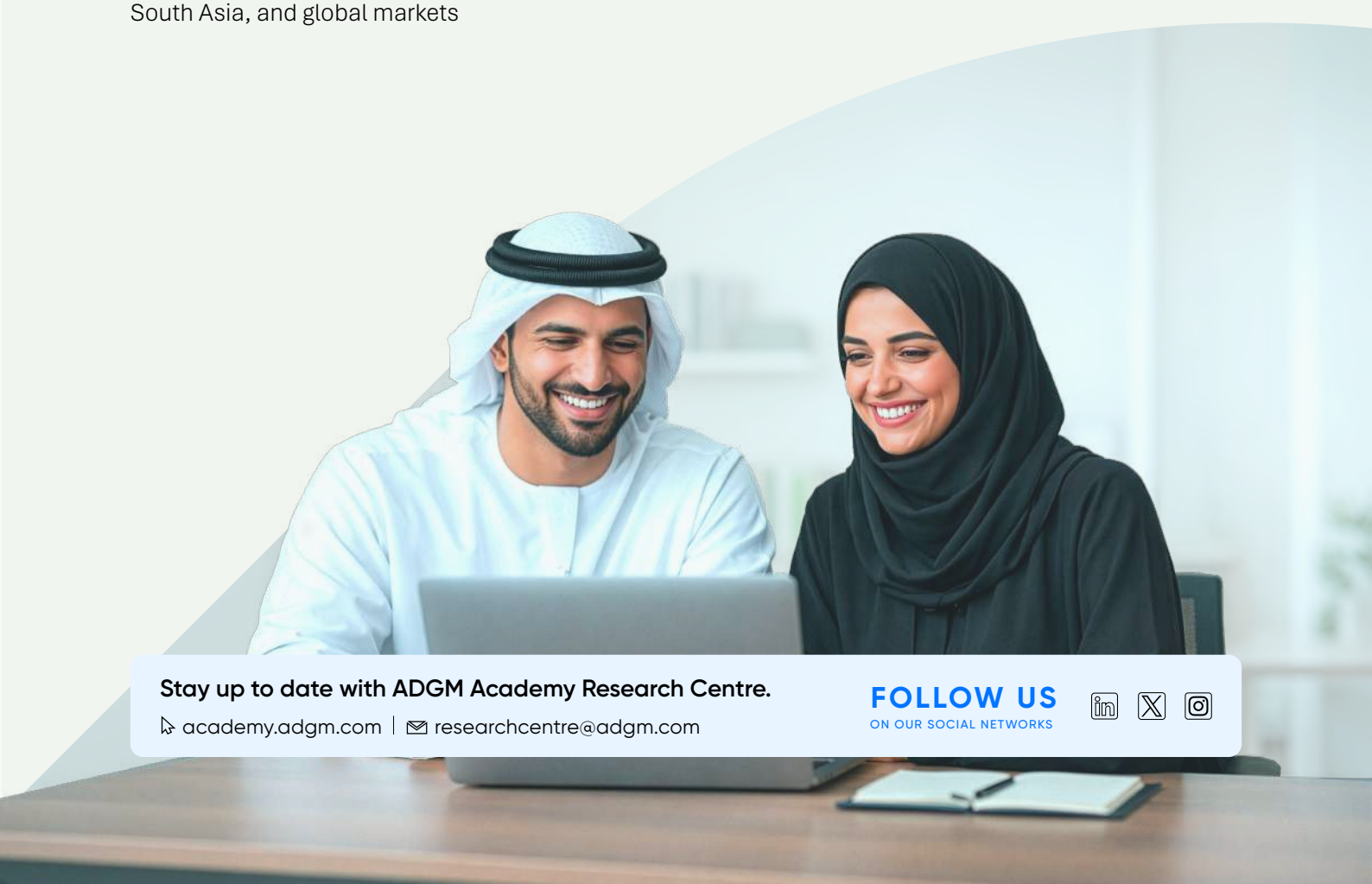
The ADGM Academy Research Centre, part of ADGM Academy, unites academics, financial practitioners, government, and technology experts to drive innovation and enhance the financial landscape in the UAE, MENA region, and beyond. As the financial sector evolves with new technologies, disruptors, and opportunities, independent research is vital to harness these changes for the benefit of businesses, customers, and society. Through collaborative insights with the academic community, the Research Centre delivers the expertise needed to navigate and capitalise on this dynamic transformation.

ADGM Academy

ADGM Academy, the knowledge hub of Abu Dhabi Global Market (ADGM), is shaping the future of banking, finance, digital innovation, and public services in the region. Committed to aligning with the UAE's vision for economic leadership, we deliver cutting-edge experiential programmes that empower both graduates and professionals and drive industry growth. As a trailblazer in financial and digital training, we collaborate with top industry experts, leading professional organizations, and renowned academic institutions to create innovative, certified programmes. Join us on a transformative journey where world-class education meets opportunity, paving the way for a stronger, smarter financial industry.

Abu Dhabi Global Market (ADGM)

ADGM is a globally recognised international financial centre that brings unique value to the emerging economy of Abu Dhabi and the broader region. Established in 2015, ADGM has significantly enhanced Abu Dhabi's stature as a leading financial centre and business hub, bolstering its role as a key player in the Falcon Economy. It serves as a vital strategic link between the growing economies of the Middle East, Africa, South Asia, and global markets



Stay up to date with ADGM Academy Research Centre.

academy.adgm.com | researchcentre@adgm.com

FOLLOW US
ON OUR SOCIAL NETWORKS

Level 20, Al Maqam Tower, ADGM Square, Al Maryah Island, PO Box 111999 Abu Dhabi, UAE
الطابـق 20، بـرج المقام، مربعة أبوظبي العالمي، جزيرة الماريـه، ص ب 111999، أبوظبي، الإمارات العربية المتحدة

academy.adgm.com