Global AI & Agentic AI Governance Landscape: A Single Unified View

A Comprehensive Reference for Practitioners, Regulators and Policy Makers

April 2026

Dr. David R. Hardoon

davidrh@me.com
www.davidroihardoon.com

Table of Contents

1. Executive Summary
2. The Urgent Need for Guidance, Governance and Regulatory Guardrails
3. From Scarcity to Overflow — The Risk of Complexity and Confusion
4. Why Exactly Three Layers — And Why These Specific Three
5. The Critical Missing Dimension: A Unified Socio-Technical System View
6. Detailed Mapping of the Global Universe into the Three Layers
7. The Unified Governance Reference — Why, What, How + SAFE
8. Practical Implementation Roadmap for Governance Practitioners
9. Worked Examples
10. Policy & Supervisory Guidance for Regulators and Policy Makers
11. Integration with Existing Standards
12. Future-Proofing the Framework
13. Conclusion and Call for Collaboration

Appendices

A. Glossary
B. Practical Artefacts and Templates
C. Decision Tree for Framework Selection
D. Curated List of Major AI Governance Initiatives
E. Visual Diagrams

How to Use This Document (Quick Navigation Guide)

- **Board / Executive**: Executive Summary + Section 7 + Section 9
- **Governance Practitioner / Risk Officer**: Sections 5, 6, 8 + Appendices B & C
- **Regulator / Policy Maker**: Sections 4, 5, 10 + Appendix D
- **Technologist / Implementer**: Sections 6, 8 + Appendices B & E

1. Executive Summary

Agentic AI has rapidly evolved from experimental pilots into core operational infrastructure. These systems now autonomously plan, invoke tools, adapt in real time, coordinate with other agents, and execute complex actions that directly impact real-world outcomes across finance, supply chains, healthcare, and enterprise operations. The opportunities are immense. Yet the risks are equally material: erroneous autonomous actions, emergent multi-agent behaviours, loss of accountability, and potential systemic instability.

The global AI governance landscape has shifted from scarcity to overwhelming abundance. The OECD AI Policy Observatory tracks more than 900 distinct national AI policies and initiatives across more than 80 jurisdictions, alongside hundreds of additional standards, practitioner playbooks, technical threat models, and sector-specific guidance. While this reflects strong global commitment, it has also created significant fragmentation and decision fatigue. Boards, risk officers, and regulators increasingly find themselves asking "Which framework do I follow?" instead of "How do I govern effectively?"

This whitepaper offers a clear, unified reference map of the global AI governance landscape. It organises the many existing frameworks and initiatives into three intuitive layers: **Why** (ethical foundation), **What** (regulatory obligations and standards), and **How** (operational playbooks and technical controls).

It anchors these layers in the Stability-Assured Framework for Entities (SAFE), a comprehensive, timeless, agent-agnostic, closed-loop engineering foundation applicable to any controller, whether human, rule-based, LLM-based, or future architectures.

SAFE models any controller as part of the same feedback loop and scores the loop using five timeless engineering metrics: Observability, Controllability, Stability, Robustness, and Performance. It produces clear Loop Risk Profiles and Autonomy Levels 0-5, enabling organisations and supervisors to shift from qualitative checklists to quantitative, auditable, and regulator-ready governance.

This unified view delivers immediate, practical value:

- **For practitioners**: a clear implementation roadmap, decision trees, templates, and worked examples that can be applied immediately.
- **For regulators and policy makers**: supervisory language, proportionality principles, and consistent measurable metrics that support examinations, policy updates, and cross-border supervisory convergence.
- **For all stakeholders:** a unified socio-technical foundation that makes governance measurable, scalable, and truly future-proof.

This whitepaper aims to be the single landing reference and authoritative map for the global AI governance community. It equips boards, risk teams, regulators, policy makers, and technologists to move from fragmentation and uncertainty to confident, coherent, and effective governance of agentic AI.

2. The Urgent Need for Guidance, Governance and Regulatory Guardrails

Agentic AI has crossed a decisive threshold. What began as experimental technology is now embedded in core operational infrastructure across every sector. These systems no longer simply predict outcomes. They autonomously plan multi-step workflows, invoke tools, adapt in real time, coordinate with other agents, and execute actions that directly shape real-world results, from executing financial transactions and optimising supply chains to coordinating patient care and driving enterprise-wide automation.

This evolution brings transformative opportunity. It also introduces qualitatively new and material risks that traditional governance approaches were never designed to address:

- Erroneous or unauthorised autonomous actions that can trigger immediate financial loss, data breaches, or patient-safety incidents.
- Emergent systemic behaviours in multi-agent fleets, such as bidding wars, herding on stale data, or uncontrolled apology cascades.
- Loss of traceability and human accountability when complex decisions arise from dynamic interactions rather than fixed rules.
- Heightened security and privacy vulnerabilities amplified by tool use, persistent memory, and inter-agent communication.

For governance practitioners, the daily reality is clear: hybrid human-AI workflows are already in production, auditors and regulators are asking hard questions, and existing frameworks treat AI in isolation rather than as part of a larger socio-technical system. For regulators and policy makers, the challenge is equally urgent: ensuring consumer protection, market stability, and supervisory consistency in an environment where agents can act faster and more widely than traditional controls can monitor.

The fundamental requirement is therefore clear. Governance must enable the adequate, proper, and sustainable integration of AI into the human operating model. This requires moving beyond AI-centric approaches and instead governing the full socio-technical system, where humans, rule-based logic, legacy systems, and AI agents function as interchangeable controllers operating within shared feedback loops.

Without a unified, coherent governance view, organisations face a stark choice between regulatory non-compliance and operational paralysis. With such a view, they can achieve confident, measurable, scalable, and regulator-ready deployment of agentic AI.

3. From Scarcity to Overflow — The Risk of Complexity and Confusion

Only a few years ago, in 2018–2023, AI governance was scarce. A handful of high-level principles, such as the OECD AI Principles and MAS FEAT, represented the main global references. By late 2025 the situation had reversed dramatically. As of April 2026 the OECD AI Policy Observatory tracks more than 900 distinct national AI policies and initiatives across more than 80 jurisdictions, complemented by hundreds of additional standards, practitioner playbooks, technical threat models, industry reports and sector-specific guidance.

This rapid proliferation reflects genuine global commitment to responsible AI. Yet it has also created a new and growing problem: fragmentation. Frameworks overlap in scope, differ sharply in emphasis, depth, enforceability and practicality. Boards, risk officers, regulators and technologists now spend significant time trying to understand which documents apply, how they relate, and which ones to prioritise. The result is not clarity but widespread decision fatigue, inconsistent internal policies, audit fatigue and the very real risk of regulatory arbitrage.

Adding yet another standalone framework would only make the situation worse. What the ecosystem urgently needs instead is synthesis and unification: a single, coherent reference map that shows exactly where every major initiative fits, how they relate to one another, and how they can be combined efficiently without duplication or conflict.

4. Why Exactly Three Layers — And Why These Specific Three

Governance in any mature, high-stakes domain, whether finance, aviation, nuclear energy or pharmaceuticals, naturally organises into three logical and practical layers. This is not an arbitrary choice. It mirrors how real-world decisions are actually made and executed in complex organisations.

- **Why** — The purpose, values and ethical foundation (the "why" we govern). This is where senior leaders and boards begin.
- **What** — The obligations, risk classifications and standards that define what must be achieved (the "what" must be done). This is the domain of regulators and compliance teams.
- **How** — The operational playbooks, processes, tools and controls that show how to achieve it in practice (the "how" to do it). This is where practitioners and engineers operate.

These three layers provide the simplest, most intuitive and most actionable way to organise the entire universe of AI governance. Using more than three layers creates unnecessary complexity and overlap. Using fewer collapses critical distinctions, for example by blurring the line between regulatory compliance and day-to-day execution. The Why–What–How structure therefore delivers clarity to every stakeholder, from board members setting strategic direction to risk officers implementing controls and regulators assessing compliance, while remaining flexible enough to accommodate future developments.

This clean layering forms the backbone of the unified view presented in this document.

5. The Critical Missing Dimension: A Unified Socio-Technical System View

A recurring and fundamental limitation cuts across virtually every existing AI governance framework: it treats AI in isolation. Most focus narrowly on the model, the agent, the prompt, the tool call or the output, while paying little attention to the full socio-technical system in which that AI actually operates.

In reality, every consequential decision-making loop involves multiple controllers working together: humans exercising judgment and bearing ultimate accountability, rule-based scripts applying deterministic logic, legacy systems, and AI agents. These controllers constantly interact through shared data, feedback loops and emergent behaviours. Treating AI alone creates dangerous blind spots in accountability, stability, controllability and long-term alignment. It also makes it impossible to integrate AI seamlessly into the human operating model at scale.

For governance practitioners this translates into daily challenges: hybrid human-AI workflows that are already in production, unclear responsibility chains during audits, and difficulty demonstrating control when agents act autonomously. For regulators and policy makers the stakes are equally high: ensuring systemic stability, consumer protection and supervisory visibility in environments where agents can act faster and more widely than traditional controls can monitor.

A true governance framework must therefore adopt a unified socio-technical system view, one that models human, rule-based and AI controllers as interchangeable parts of the same closed feedback loop. This is the only reliable way to manage emergent risks at fleet scale, maintain human accountability when agents act autonomously, and ensure long-term stability and alignment.

SAFE (Stability-Assured Framework for Entities, Hardoon, November 2025) is the only known framework that provides this complete, agent-agnostic, closed-loop engineering foundation. It models any controller, whether human, rule-based, LLM-based or future architectures, as part of the same feedback system and evaluates the loop itself using five timeless engineering metrics. SAFE therefore serves as the theoretical baseline and engineering foundation for the entire governance universe. It underpins and strengthens all three layers while remaining fully compatible with them.

6. Detailed Mapping of the Global Universe into the Three Layers

The global AI governance landscape is now mature enough to be organised clearly. All major frameworks fit naturally into the three layers introduced earlier. The mapping below shows the most influential initiatives in each layer, how they relate to one another, their key benefits and gaps, and practical guidance on which ones to prioritise.

Why Layer – Foundational Principles and Ethical Values

Key frameworks (selected from the 900+ tracked) OECD AI Principles, UNESCO Recommendation on the Ethics of AI, MAS FEAT Principles, G7 Hiroshima AI Process, G20 AI Principles, Council of Europe Framework Convention on Artificial Intelligence.

How they relate Later documents explicitly reference and build upon earlier ones and converge on the same core values: fairness, transparency, accountability, human-centricity, robustness and sustainability.

Benefits and gaps Global moral consensus and a common language, but purely aspirational with no metrics or enforcement.

Guiding principle – Which one to follow? Start with the OECD AI Principles as the primary global reference. Supplement with MAS FEAT if you operate in financial services or Singapore-influenced jurisdictions.

Unified meta-framework A single Ethical Reference Model that defines high-level values and desired outcomes. These values become the explicit Reference Input in any control loop.

What Layer – Regulatory Obligations and Standards

Key frameworks EU AI Act, NIST AI RMF (including 2026 Agent Update), ISO/IEC 42001, China Artificial Intelligence Safety Governance Framework 2.0, South Korea Framework Act on AI, Singapore MAS AI Guidelines, US Executive Orders and state laws, UK AI Regulation White Paper updates.

How they relate The EU AI Act has become the de-facto global benchmark. Many national laws reference or align with it, while standards such as ISO and NIST provide certifiable overlays.

Benefits and gaps Enforceable obligations, risk classification and auditability, but still largely static and model-centric with limited treatment of dynamic trajectories or multi-agent emergence.

Guiding principle – Which one to follow?

- Operating in or serving Europe: treat the EU AI Act as the baseline.
- Global operations: align to NIST AI RMF and ISO/IEC 42001 for certification.
- Financial services: layer MAS AI Guidelines on top. Always adopt the strictest applicable regulation as the baseline and map others to it.

Unified meta-framework A single Risk-Tier + Obligation Model that translates ethical principles into enforceable requirements and risk classifications.

How Layer – Operational Playbooks and Technical Controls

Key frameworks IMDA Model AI Governance Framework for Agentic AI, Szpruch et al. (Scalable Runtime Governance for Agentic AI in Financial Services), OWASP Agentic AI Threats & Mitigations, CSA Draft Addendum on Securing Agentic AI, WEF AI Agents in Action, and major consultancy playbooks from McKinsey, Deloitte, EY, PwC, BCG and KPMG.

How they relate Practitioner playbooks such as IMDA and Szpruch translate regulations into actionable steps. Technical papers such as OWASP and CSA add specific controls and threat models on top.

Benefits and gaps Concrete checklists, testing methods, monitoring guidance and implementation artefacts, but mostly qualitative or sector-specific and lacking universal quantitative metrics and true agent-agnosticism.

Guiding principle – Which one to follow?

- General organisations: start with IMDA MGF.
- Financial services or regulated banking: use Szpruch et al. as the primary MRM playbook.
- Technical or security teams: layer OWASP and CSA on top.
- Need for multi-agent fleet governance: combine with SAFE's hierarchical and dissipativity tools.

Unified meta-framework A single Operational Implementation Model covering design limits, capability decomposition, runtime monitoring, testing, human oversight and continuous governance.

High-Level Comparison of Influential Frameworks Subset

Framework	Layer	Primary Strength	Key Gap	Best Used For	Relation to SAFE
OECD AI Principles	Why	Global moral consensus and common language	No metrics or enforcement	Strategic direction and board alignment	Provides the Reference Input for loops
UNESCO Ethics of AI	Why	Broad ethical foundation	Purely aspirational	International policy alignment	Supplies ethical values as measurable references
MAS FEAT Principles	Why	Finance-sector relevance	Limited to principles	Financial services strategy	Strong foundation for financial use cases

Framework	Layer	Primary Strength	Key Gap	Best Used For	Relation to SAFE
EU AI Act	What	Enforceable risk-based obligations	Static and model-centric	EU operations or EU customers	Risk tiers map directly to Autonomy Levels
NIST AI RMF (2026 update)	What	Flexible, widely respected	Limited runtime focus	Global operations and US alignment	Excellent overlay for SAFE metrics
ISO/IEC 42001:2023	What	Certifiable management system	Static, not agent-dynamic	Certification and audit readiness	Maps to observability and controllability clauses
IMDA MGF for Agentic AI	How	Practical four-pillar checklist	Qualitative, not quantitative	General enterprise deployment	Provides implementation steps for SAFE metrics
Szpruch et al. (2026)	How	Banking-specific runtime MRM playbook	Finance-sector only	Regulated financial institutions	Strong operational delivery layer for SAFE
OWASP Agentic AI Threats	How	Concrete threat mitigations	Reactive, not systemic	Security and technical teams	Maps directly to SAFE Robustness metric
WEF AI Agents in Action	How	Evaluation and governance foundations	High-level	Cross-industry evaluation	Useful complement for SAFE Performance

Table 1: High-level comparison of the most influential frameworks. SAFE serves as the cross-cutting engineering foundation that supplies measurable metrics to every layer.

The three layers together provide a complete picture. The Why layer sets direction, the What layer sets obligations, and the How layer delivers execution. SAFE sits across all three as the engineering foundation that makes every layer measurable and actionable.

7. The Unified Governance Reference — Why, What, How + SAFE as Theoretical Baseline

The three layers provide the structure. SAFE provides the foundation that makes the entire structure work.

SAFE sits as the cross-cutting Theoretical Baseline and Engineering Foundation. It does not replace the Why, What or How layers. Instead, it unifies and strengthens all three by supplying the missing universal, quantitative, agent-agnostic language and closed-loop model that turns high-level principles and obligations into measurable, auditable governance.

The Complete Unified Meta-Framework

Why Layer – Ethical Reference Model All principles become the high-level Reference Input in the control loop. SAFE contribution: Turns values into explicit, version-controlled, measurable references.

What Layer – Risk-Tier + Obligation Model All regulations and standards become required metric thresholds and Autonomy Levels 0-5. SAFE contribution: Provides the quantitative proof of compliance (Loop Risk Profile).

How Layer – Operational Implementation Model All playbooks and technical controls become the machinery that measures and enforces SAFE's five metrics. IMDA's four pillars become concrete steps for achieving SAFE metrics. Szpruch's capabilities and telemetry become the banking-grade instrumentation substrate. OWASP and CSA threats map to robustness and controllability gaps.

SAFE – Theoretical Baseline / Engineering Foundation (cross-cutting) The single closed-loop model that works for any controller (human, rule-based, LLM or future). It supplies the five timeless metrics, Loop Risk Profile, Autonomy Levels 0-5, hierarchical supervisory control, dissipativity bounds and robust reference design.

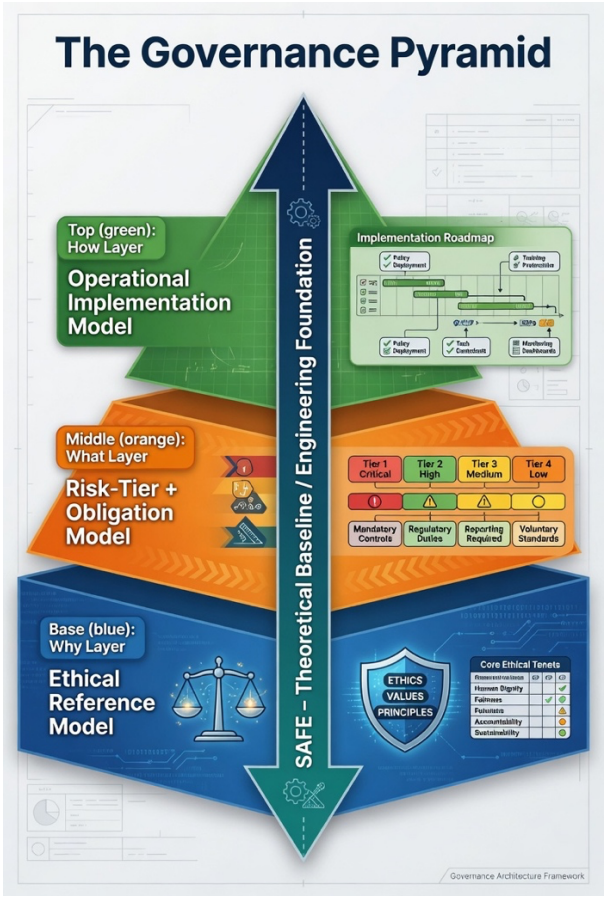

Diagram 1: The Governance Pyramid

Diagram 2: Ecosystem Map – All Frameworks Orbiting SAFE

8. Practical Implementation Roadmap for Governance Practitioners

This roadmap translates the unified Why–What–How + SAFE framework into concrete, phased action. It is designed for immediate use by governance, risk, and compliance teams.

Phase 1: Foundation (Weeks 1–4) Map all existing and planned AI/agentic systems to SAFE's closed-loop model. Score each loop against the five metrics using the SAFE scorecard template (Appendix B). Align organisational values to the Why layer (OECD AI Principles as baseline, supplemented by MAS FEAT where relevant).

Phase 2: Compliance & Risk Tiering (Weeks 5–8) Map applicable regulatory obligations from the What layer to SAFE Autonomy Levels and required metric thresholds. Perform a gap analysis against ISO/IEC 42001 and any sector-specific rules. Produce the first organisation-wide Loop Risk Profile report.

Phase 3: Operational Controls (Months 2–3) Select the primary How-layer playbook: IMDA MGF for general enterprise use or Szpruch et al. for regulated financial services. Implement capability catalogues, governance-semantic telemetry, deterministic guards, and human oversight checkpoints. Deploy the chosen playbook on top of the SAFE metrics.

Phase 4: Continuous Governance & Maturity (Month 4 onward) Establish quarterly Loop Risk Profile reviews and automated monitoring. Integrate findings into existing 3LoD and MRM processes. Monitor multi-agent emergent risks using SAFE's hierarchical supervisory control and dissipativity bounds. Drive continuous improvement toward Maturity Level 4 (Unified & Optimised).

Maturity Model

- Level 1: Ad-hoc – AI treated in isolation.
- Level 2: Layered – Why, What and How adopted separately.
- Level 3: SAFE-grounded – Five metrics and Loop Risk Profiles actively used.
- Level 4: Unified & Optimised – Full socio-technical integration with continuous improvement.

Sample Deliverable: SAFE Scorecard Template (excerpt – full version in Appendix B)

Property	Score (1–5)	Evidence / Metric	Mitigation Required?	Target
Observability		% decisions with full CoT logging & provenance (<5 s reconstruction)		≥4.0
Controllability		Override latency (99th percentile) + privilege boundaries		≥4.0
Stability		Worst-case overshoot in red-team scenarios		≥4.0
Robustness		Max tolerated adversarial input		≥4.0
Performance		Median settling time vs reference KPI		≥4.0

9. Worked Examples

The following worked examples illustrate how the unified Why–What–How + SAFE framework is applied in practice. Each case shows the real-world context, the specific steps taken across the layers, the concrete application of SAFE's metrics and tools, and the measurable outcomes achieved.

Example 1: Banking Credit-Memo Agent (Regulated Financial Services – Szpruch + SAFE) A Tier-1 bank introduced an agentic AI to draft credit memos for corporate lending. Initially the agent operated with only basic prompt engineering and human review (a pure How-layer approach). This led to inconsistent quality, long review cycles, and growing audit concerns about traceability.

The bank adopted the unified view as follows:

- **Why layer**: Aligned the agent's reference inputs to the bank's credit policy and OECD/MAS FEAT ethical values.
- **What layer**: Mapped the use case to EU AI Act high-risk obligations and NIST RMF requirements, setting a target Autonomy Level of 4.
- **How layer**: Implemented Szpruch's capability decomposition (defining the credit-memo drafting capability with explicit authority, constraints, and evidence packs) and execution trajectories with governance-semantic telemetry.
- **SAFE foundation**: Scored the full loop on the five metrics (Observability = 4.2 with full chain-of-thought logging and provenance; Controllability = 4.5 with hierarchical approval gates and override latency under 2 seconds; Stability, Robustness, and Performance all above 4.0).

The resulting Loop Risk Profile of 4.3 justified Autonomy Level 4 with targeted human monitoring. Deterministic guards blocked any output outside policy parameters. **Outcome**: Manual review time dropped by 65 %, first-quarter audit findings fell to zero, and the 2LoD team received clear, regulator-ready evidence of control. The bank now uses the same pattern for other lending workflows.

Example 2: Enterprise Multi-Agent Procurement Fleet (Cross-Industry – IMDA + SAFE) A global manufacturing company deployed four coordinated procurement agents to source raw materials. Using only IMDA's design limits and human oversight checkpoints, the fleet occasionally triggered bidding wars on stale supplier data, leading to cost overruns and supplier disputes.

Application of the unified framework:

- **Why layer**: Reference inputs were set using the company's sustainability and fair-procurement values (OECD principles).
- **What layer**: The use case was assessed under ISO 42001 and relevant national procurement regulations, targeting Autonomy Level 3.
- **How layer**: IMDA's four-pillar checklist provided the operational playbook (design limits on tool access, human oversight checkpoints).
- **SAFE foundation**: The fleet was modelled as a multi-agent closed loop. SAFE's dissipativity bounds and hierarchical supervisory controller were added to prevent

emergent instability. Real-time scoring showed fleet-level stability rising from 2.1 to 4.6 after implementation.

Outcome: Procurement cycle time improved by 40 %, unintended over-spending was eliminated, and a single quarterly Loop Risk Profile report satisfied both internal governance and external auditors. The company has since scaled the same pattern to logistics and vendor-management fleets.

Example 3: Hybrid Human–AI Customer Service Loop (Regulated Telecom Sector – SAFE Core) A major telecom provider combined human agents with AI escalation agents for complex billing and technical queries. Early deployments suffered from unclear escalation triggers and accountability gaps during regulatory reviews.

Framework application:

- **Why layer**: Ethical reference inputs were drawn from customer-centric values and MAS FEAT principles.
- **What layer**: The use case was classified under relevant consumer-protection rules and assigned a target Autonomy Level of 3.
- **How layer**: Human oversight checkpoints and escalation protocols were defined.
- **SAFE foundation**: Both human and AI agents were modelled as interchangeable controllers in the same closed loop. Observability and controllability metrics triggered automatic escalation to humans when confidence dropped below threshold. The full loop was scored quarterly, ensuring consistent accountability.

Outcome: 82 % of queries resolved autonomously while maintaining full audit trails. Regulators accepted the Loop Risk Profile as sufficient evidence of proportionate oversight, reducing examination findings by 70 % and shortening response times for customers.

These examples demonstrate that the unified view is not theoretical. When the three layers are grounded in SAFE's engineering foundation, governance becomes measurable, scalable, regulator-ready, and directly applicable across industries and risk levels. Organisations can replicate the same pattern with confidence, knowing exactly which tools from each layer to apply and how SAFE quantifies success.

10. Policy & Supervisory Guidance for Regulators and Policy Makers

Regulators and supervisors face a shared challenge: how to oversee agentic AI systems consistently and proportionately without creating additional complexity or regulatory burden. Existing frameworks provide strong principles and obligations, yet many supervisors still struggle with three practical gaps:

- Measuring the effectiveness of controls in real time rather than relying on documentation alone.
- Demonstrating clear accountability in hybrid human-AI and multi-agent environments.
- Achieving supervisory convergence across jurisdictions and sectors while preserving flexibility.

The unified Why–What–How + SAFE framework is designed to close these gaps directly. It does not replace existing regimes; it strengthens them by providing a single, timeless, quantitative engineering foundation that supervisors can reference without introducing new processes.

How Supervisors Can Apply the Framework

- Integrate into existing risk-based supervision using SAFE's five metrics and Loop Risk Profile as a practical way to assess the quality of an institution's controls.
- Support proportionality with Autonomy Levels 0–5 as a clear graduated scale.
- Enhance cross-border and cross-sector consistency using the same SAFE metrics as a common benchmark.

Suggested Flexible Language for Guidance or Examination Manuals

Supervisors may choose to include wording such as:

"Institutions should be able to demonstrate that high-impact agentic AI systems are governed through a closed-loop approach that is scored against objective metrics of observability, controllability, stability, robustness, and performance, resulting in a defined autonomy level. This information should be maintained and made available upon supervisory request."

Example Supervisory Finding / Examination Question Sample Examination Question: *"Provide the most recent Loop Risk Profile and Autonomy Level for each high-impact agentic AI system, together with evidence of the five SAFE metrics and any mitigating controls applied."*

Sample Supervisory Finding: *"The institution has demonstrated effective governance of its credit-memo agent through a SAFE Loop Risk Profile of 4.3 (Autonomy Level 4). Observability and controllability metrics meet supervisory expectations; continued quarterly reporting is required."*

Practical Benefits for Supervisors

- Provides clearer evidence during examinations.
- Reduces reliance on narrative descriptions by offering comparable quantitative indicators.

- Supports proportionality and risk-based supervision without adding layers of process.
- Facilitates dialogue with institutions and across borders using a common technical vocabulary.

SAFE does not create new obligations. It equips supervisors with a practical, measurable way to address the real gaps they already encounter when overseeing agentic AI.

11. Integration with Existing Standards

The unified Why–What–How + SAFE framework is deliberately designed to complement and strengthen existing standards and regulatory regimes rather than replace them. SAFE supplies the quantitative, measurable engineering layer that many current frameworks currently lack, making compliance easier to demonstrate and supervise.

Key Integration Points

- **ISO/IEC 42001:2023 (AI Management System)** SAFE's five metrics directly operationalise clause A.6 (monitoring and measurement) and clause A.5 (human oversight). Institutions can use the Loop Risk Profile and supporting telemetry as auditable evidence during certification audits.

- **Model Risk Management Programmes (SR 11-7, PRA SS1/23, MAS MRM, and the April 17, 2026 Revised Interagency Guidance)** The April 17, 2026 interagency revision (FDIC, OCC, Federal Reserve) updates traditional model risk management expectations for conventional models. It does not address agentic or generative AI systems. Szpruch et al. provides the specialised runtime governance artefacts (capability catalogue, evidence packs, governance-semantic telemetry) for agentic systems, while SAFE supplies the quantitative loop scoring and Autonomy Levels. Together they create a complete, regulator-ready package that bridges the gap between classical MRM and agentic AI.

- **Three Lines of Defence (3LoD)** SAFE enables clear allocation of responsibilities across the lines of defence:
    - 1LoD owns and scores the loops.
    - 2LoD challenges the metrics and risk profiles.
    - 3LoD audits the telemetry and evidence packs.

Practical Benefit Organisations can use the same SAFE scorecard and Loop Risk Profile across multiple standards and regimes. This reduces duplication of effort, simplifies internal governance processes, and provides consistent, comparable evidence for external reporting and supervisory examinations.

12. Future-Proofing the Framework

SAFE is deliberately built to remain relevant across technological generations. Its five engineering metrics and closed-loop model are architecture-agnostic and timeless. They apply equally to today's LLM-based agents, tomorrow's quantum or neuromorphic controllers, and any future hybrid systems that combine human, rule-based, and artificial decision-making.

Organisations and supervisors can therefore adopt the framework with confidence that it will not require wholesale replacement when new controller types emerge. The core closed-loop structure and five metrics stay constant. Only the proxy measures and evidence standards are updated periodically to reflect the latest technological capabilities, ensuring continued relevance without disrupting established governance processes.

This design delivers two critical assurances:

- Practitioners can invest in capability catalogues, telemetry infrastructure, and Loop Risk Profiles knowing the investment will endure.
- Regulators and policy makers gain a stable, consistent benchmark that supports long-term supervisory convergence and reduces the risk of regulatory obsolescence.

By anchoring governance in fundamental control-theoretic principles rather than any specific technology, the unified view presented in this document is built to last.

13. Conclusion and Call for Collaboration

The global AI governance landscape has reached a tipping point. From a handful of high-level principles a decade ago, we now have more than 900 distinct initiatives, standards, playbooks, and guidance documents. The challenge is no longer scarcity but fragmentation and decision fatigue.

This whitepaper offers a clear, unified reference map of the global AI governance landscape. It organises the many existing frameworks and initiatives into three intuitive layers — Why (ethical foundation), What (regulatory obligations and standards), and How (operational playbooks and technical controls) — and anchors them in the Stability-Assured Framework for Entities (SAFE), a comprehensive, timeless, agent-agnostic, closed-loop engineering foundation applicable to any controller, whether human, rule-based, LLM-based, or future architectures.

SAFE helps bring greater coherence and measurability to this complex landscape. It equips boards, risk teams, regulators, policy makers, and technologists to move from fragmentation and uncertainty toward more confident, coherent, and effective governance of agentic AI.

This living reference will be updated periodically. I welcome your feedback, case studies, translations, and collaboration. Please submit contributions at www.davidroihardoon.com.

Appendices

Appendix A: Glossary

- **Closed-loop model**: Every controller (human, rule-based, LLM or future) is modelled as part of a feedback loop with reference input, controller, actuator, plant, sensors and disturbances.
- **Five SAFE metrics**: Observability, Controllability, Stability, Robustness, Performance – scored 1–5.
- **Loop Risk Profile**: Weighted aggregate of the five metrics used to derive Autonomy Level.
- **Autonomy Level**: 0–5 scale indicating the degree of independent operation a loop may safely have.
- **Dissipativity**: Mathematical guarantee that a multi-agent fleet cannot create energy (i.e., amplify disturbances) indefinitely, ensuring fleet-level stability.
- **Capability (Szpruch et al.)**: Reusable, bounded AI component with explicit authority, constraints and evidence requirements.
- **Trajectory**: Sequence of states and actions in a labelled transition system (Szpruch et al.).

Appendix B: Practical Artefacts and Templates

SAFE Scorecard Template (full)

Property	Score (1–5)	Evidence / Metric	Mitigation Required?	Target
Observability		% decisions with full CoT logging & provenance (<5 s reconstruction)		≥4.0
Controllability		Override latency (99th percentile) + privilege boundaries		≥4.0
Stability		Worst-case overshoot in red-team scenarios		≥4.0
Robustness		Max tolerated adversarial input		≥4.0
Performance		Median settling time vs reference KPI		≥4.0

Filled Example: Banking Credit-Memo Agent (Q2 2026 review)

Property	Score (1–5)	Evidence / Metric	Mitigation Required?	Target	Actual
Observability	4.3	98 % decisions with full CoT logging & provenance (<3 s reconstruction)	No	≥4.0	Met
Controllability	4.6	Override latency 1.8 s (99th percentile) + 3 independent privilege boundaries	No	≥4.0	Met
Stability	4.1	Worst-case overshoot 6 % in 100 red-team scenarios	Minor tuning	≥4.0	Met
Robustness	4.4	Max tolerated adversarial input (AdvBench-style)	No	≥4.0	Met
Performance	4.2	Median settling time 4.2 s vs reference KPI	No	≥4.0	Met

Overall Loop Risk Profile: 4.32 → Autonomy Level 4 approved

Appendix C: Decision Tree for Framework Selection

- Start → Is the use case in regulated financial services? →
    - Yes → Primary playbook: Szpruch et al. + SAFE metrics for loop scoring. →
    - No → General enterprise? → Primary playbook: IMDA MGF + SAFE metrics for quantification.
- Need multi-agent fleet stability? → Apply SAFE hierarchical supervision and dissipativity bounds on top of chosen playbook.

Appendix D: Curated List of Major AI Governance Initiatives (80+ Most Influential)

The OECD AI Policy Observatory currently tracks more than 900 distinct national AI policies and initiatives. The complete, searchable, and regularly updated database is available at: **https://oecd.ai/en/dashboards/policy-initiatives**

Why Layer

- OECD AI Principles – https://oecd.ai/en/ai-principles
- UNESCO Recommendation on the Ethics of AI – https://www.unesco.org/en/artificial-intelligence/ethics
- MAS FEAT Principles – https://www.mas.gov.sg/publications/monographs-or-information-paper/2019/feat

What Layer

- EU AI Act – https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202401689
- NIST AI RMF 1.0 + 2026 Agent Update – https://www.nist.gov/itl/ai-risk-management-framework
- ISO/IEC 42001:2023 – https://www.iso.org/standard/81230.html

How Layer

- IMDA Model AI Governance Framework for Agentic AI (Jan 2026)
- Szpruch, Lukasz and Sudjianto, Agus and Bhatti, Tanveer and Ang, Gary, Scalable Runtime Governance for Agentic AI in Financial Services (April 13, 2026): SSRN abstract 6567199
- OWASP Agentic AI Threats & Mitigations – https://genai.owasp.org

Appendix E: Visual Diagrams

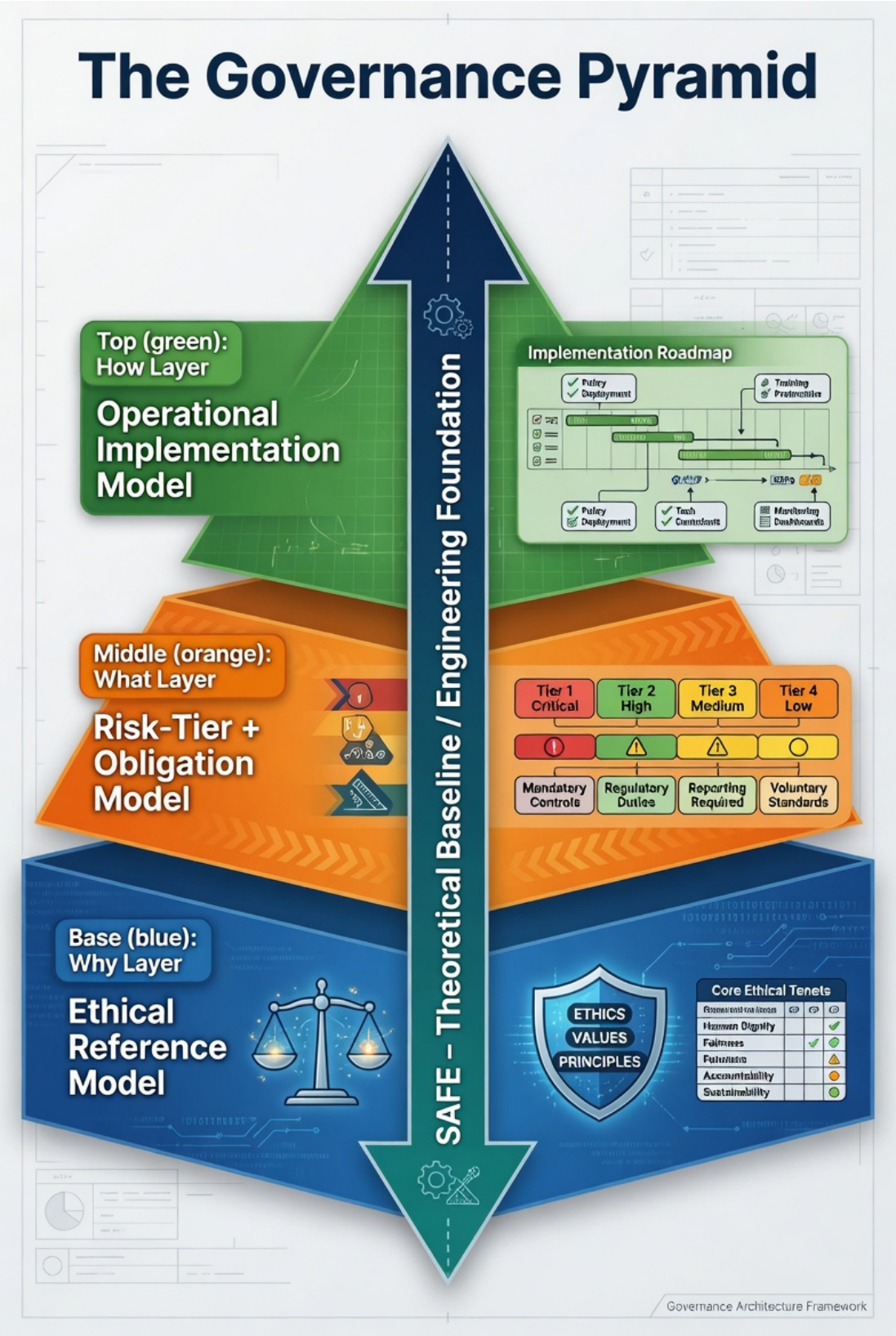

Diagram 1: The Governance Pyramid

Diagram 2: Ecosystem Map – All Frameworks Orbiting SAFE