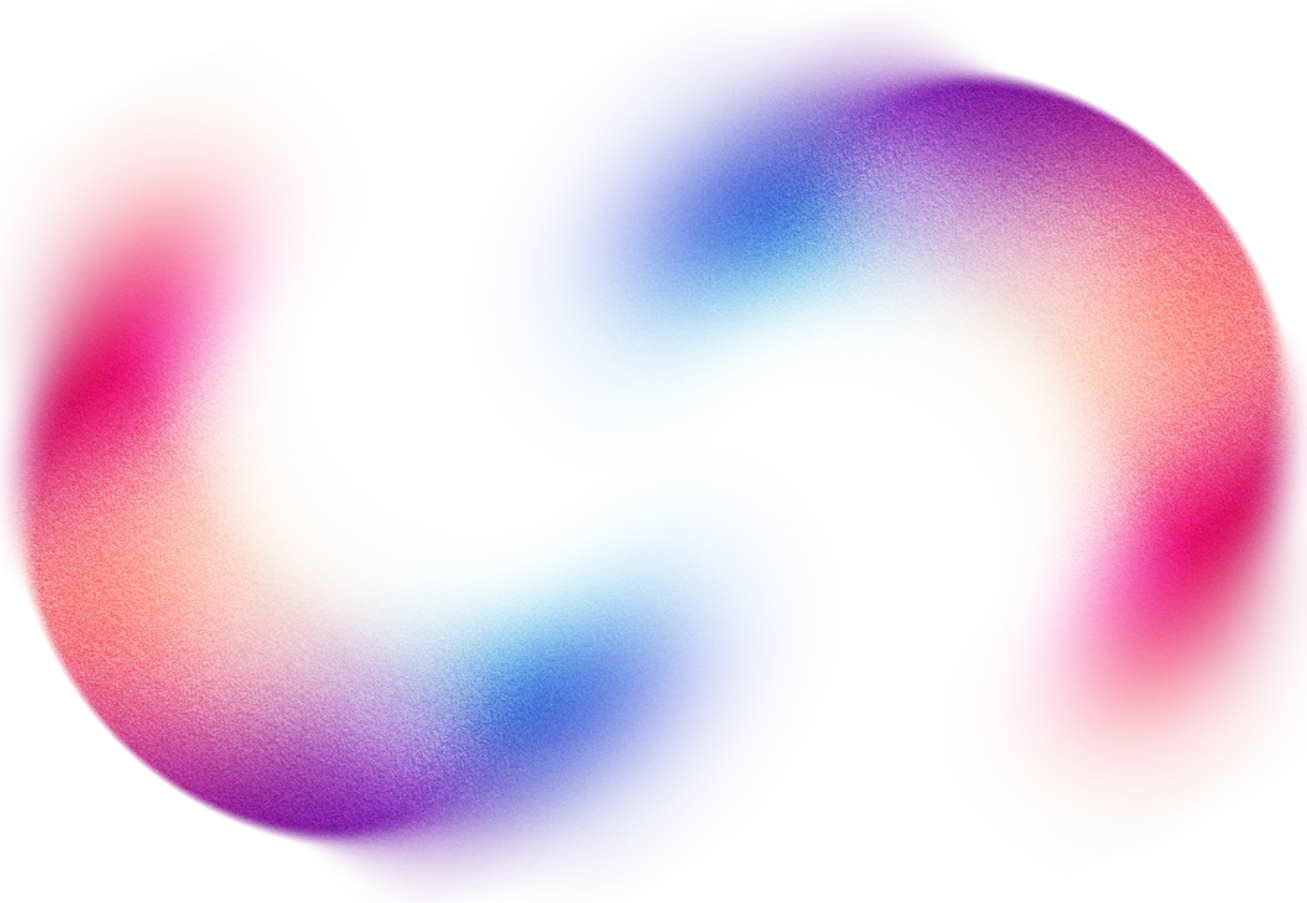
AI GOVERNANCE PAPER SERIES

MyConsultancy

AI RISK, CONTROL AND ASSURANCE AT SCALE

Governing AI Where It Actually Behaves



Manoj Tavarajoo

Executive Summary

AI risk is not a problem of intent. Most organisations that have experienced material AI failures had governance frameworks in place. They had policies, approval processes and documented controls. What they did not have was assurance that those controls were operating in production, when and where it mattered.

Recent industry analysis suggests that over 70% of organisations now have generative AI in production, while only a small minority have established mature assurance capabilities to monitor and control those systems in operation. The gap between deployment and governance is not closing. It is widening.

This paper argues that AI risk cannot be managed through approval-based governance alone. It must be controlled and assured continuously across the full lifecycle of AI systems, from intake and design through deployment, operation and ongoing use. Controls must be embedded in how AI systems run, not appended as documentation.

> The organisations that govern AI effectively are not those that approve use cases fastest. They are those that can demonstrate, through evidence, that their AI systems are behaving safely, reliably and within defined boundaries at any point in time.

The Governance Problem

In recent years, regulators have repeatedly identified failures in AI-driven decision systems that had already passed internal governance processes. In financial services, automated decision models have produced discriminatory outcomes despite formal approval and validation. In customer service, generative AI systems have produced incorrect or misleading responses that were not detected until customers surfaced them externally.

These are not isolated incidents. They reflect a consistent pattern.

AI governance has matured. Organisations have introduced responsible AI frameworks, ethics principles, review boards and approval gates. Board attention has increased. Regulatory expectations have intensified. Yet the quality of assurance in production environments remains inconsistent.

> The gap is structural. Governance is still concentrated at the point of approval. AI risk, however, does not peak at approval. It accumulates in operation.

Why AI Risk Grows as AI Scales

Traditional systems are largely deterministic. Once deployed, they behave predictably within defined parameters. Risk is concentrated in design and release.

AI systems behave differently. They are probabilistic and data dependent. A model trained on historical data operates in a changing environment. Input data shifts, user behaviour evolves and external conditions change. Performance can degrade gradually and without visible signals until the consequences become material.

As organisations scale AI, two shifts occur simultaneously. First, the number of models in production increases, each carrying its own risk profile. Second, AI becomes embedded in decision-making at scale. A single model does not produce one outcome. It produces thousands of outcomes across customers, transactions and operations. Small deviations in behaviour scale into material consequences. A model producing subtly biased outputs does not create one problem. It creates problems at the volume of every transaction it influences.

The National Institute of Standards and Technology AI Risk Management Framework formalises this by defining risk management as a continuous lifecycle across Govern, Map, Measure and Manage functions. Risk is not static. It must be managed continuously, not assessed once and assumed stable.

> AI risk scales faster than traditional governance mechanisms were designed to handle.

The False Confidence of Compliance-Only Governance

The most common failure mode in AI governance is not the absence of controls. It is misplaced confidence in controls that are not operating as assumed.

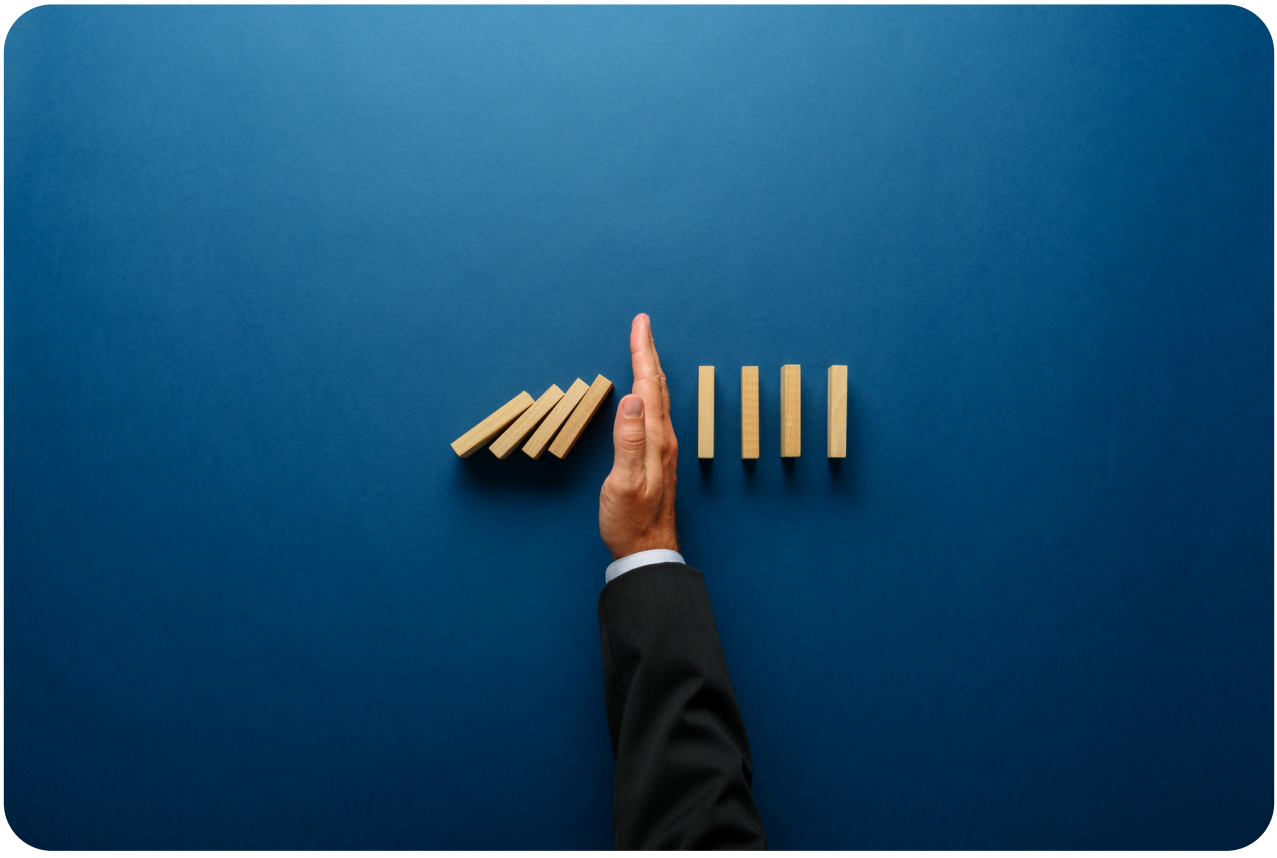
Organisations with mature governance frameworks typically have strong documentation. They maintain use case inventories, model documentation, approval records and accountability structures. Governance forums operate regularly. Reporting is produced. Executives receive dashboards summarising AI activity.

Yet evidence of control effectiveness in production environments is often weak or entirely absent. Approval-based governance confirms that a decision was made. It does not confirm that the controls underpinning that decision are still functioning months later in a changed data environment.

The EU AI Act reinforces this distinction. For high-risk AI systems, the Act requires post-deployment monitoring, incident management and ongoing oversight. Compliance is not achieved through documentation of intent. It requires operational evidence of control.

Recent updates to interagency model risk management guidance from US regulators, including the OCC, Federal Reserve and FDIC in 2026, have explicitly highlighted that generative AI and emerging agentic systems do not fit within traditional model risk management frameworks. These systems introduce behaviours and risks that extend beyond the assumptions underpinning existing guidance.

> Governance that stops at approval is incomplete. Real governance must extend into operation, where risk actually materialises.



AI Risk Categories and Control Alignment

AI risk must be understood across a structured set of categories that interact and compound at scale. Strategic risk covers failure to align AI with business objectives or deliver expected value. Operational risk arises from model errors, instability and process breakdowns. Financial risk emerges from incorrect decisions, cost overruns or reporting inaccuracies. Compliance risk involves breaches of regulatory or legal obligations. Ethical risk encompasses bias, unfair outcomes and lack of explainability. Data privacy risk arises from the misuse or exposure of sensitive personal data. Security risk includes adversarial attacks, prompt injection and model manipulation. Reputational risk is the cumulative consequence of any of these risks materialising publicly.

These categories are not independent. A failure in data quality can become a model performance issue. A model failure can become a compliance breach. A compliance breach can become a reputational crisis. Effective governance requires that controls are aligned across all categories and that each is addressed through a combination of preventive, detective and corrective controls working together. Without that alignment, gaps emerge quickly and compound silently.

> The challenge is not identifying these risks. It is ensuring they are consistently controlled across the lifecycle of every system in production.

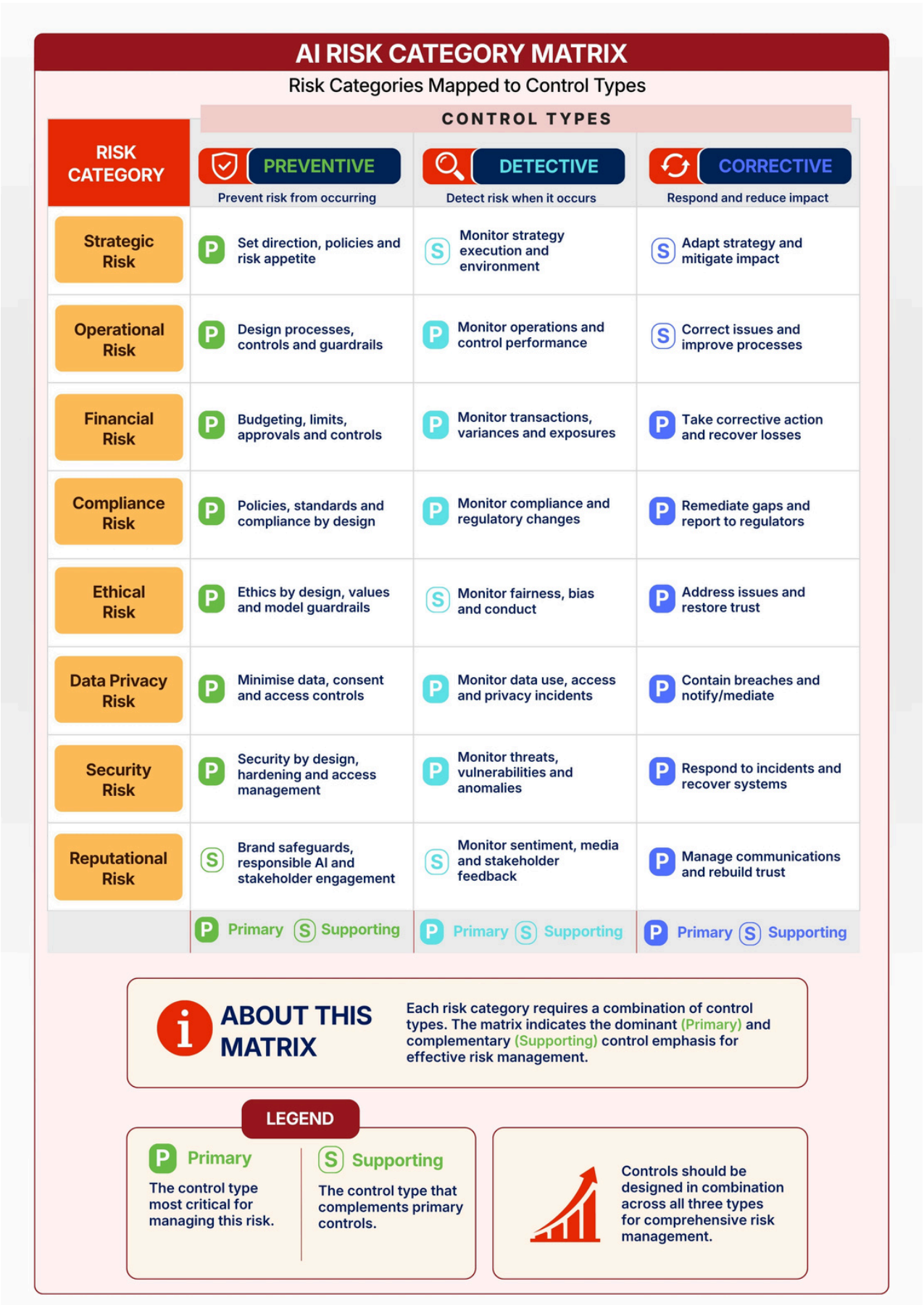


Figure 1: AI Risk Category Matrix

The AI Risk and Control Lifecycle

AI risk management must operate as a continuous lifecycle. Controls cannot be treated as a checkpoint at deployment. They must be designed, embedded, operated and improved across the full life of every AI system in production.

The lifecycle begins at intake and risk classification. Before a use case is approved for development, it must be assessed against a structured risk framework. This classification determines the control standard that will apply throughout the lifecycle. A system classified as high risk at intake will require independent validation, stronger monitoring and more rigorous assurance than a lower-risk system. Classification is not a formality. It is the governance decision that defines everything that follows.

During design and development, controls focus on data governance, model selection, fairness assessment and technical validation. These controls reduce the likelihood of problems entering production. Pre-deployment testing and validation confirm that the system performs as expected under realistic conditions. In regulated environments, independent validation by a function separate from the development team is often required.

Deployment and release controls govern how a model moves into production. Without these controls, the approved version of a model and the version operating in practice can diverge without detection. Production monitoring then becomes the critical capability, and the one where most organisations are weakest. Once a system is live, oversight frequently becomes informal. Performance drift accumulates. Threshold breaches go unaddressed. The governance present at approval is no longer visible in operation.

Incident response must be predefined, not improvised. Continuous improvement closes the cycle, with periodic revalidation ensuring systems remain appropriate for their operating environment and lessons from incidents are systematically embedded.

Controls are not a checkpoint. They are a continuously operating system embedded across the full lifecycle.

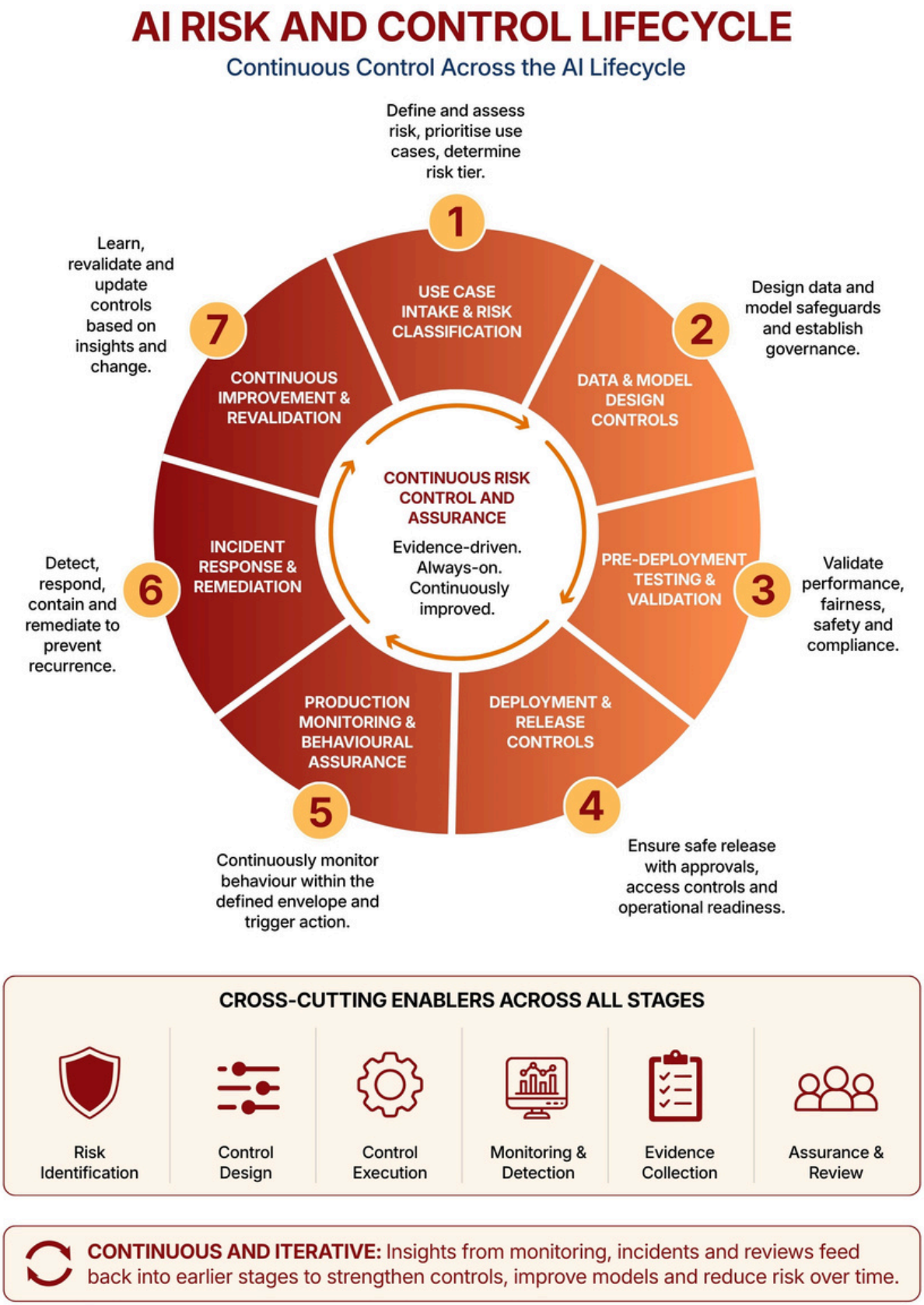


Figure 2: AI Risk and Control Lifecycle

Designing Controls That Actually Work

Effective control design requires balance across three control types that must work together.

Preventive controls reduce the likelihood of risk materialising. Detective controls increase visibility and identify problems early. Corrective controls reduce the impact of failures when they occur.

Many organisations rely disproportionately on preventive controls. Policies, approval gates and documentation requirements are preventive in nature and well suited to formal governance processes. They are necessary but insufficient. Preventive controls cannot detect problems that emerge after deployment. An organisation with strong preventive governance and weak detective capability will not know something is wrong until consequences become visible externally.

The reverse also fails. Monitoring without strong upstream governance produces alerts that cannot be acted on effectively because accountability is unclear and corrective protocols are absent.

Control design must be risk-proportionate. High-risk systems require more rigorous monitoring and more structured assurance than low-risk systems. A governance model that applies the same control standard to all AI systems will either over-govern routine use cases or under-govern consequential ones.

Controls must also be embedded in systems and workflows, not documented alongside them. A control that depends on manual application is not a control. It is an intention rather than an assurance.

Generative AI introduces an additional risk that is not present in traditional systems. Foundation model providers can update underlying models without explicit notification, altering behaviour in ways that bypass internal validation processes. This silent versioning means that a system that was previously validated can move outside acceptable performance boundaries without any internal change being made. This reinforces the need for continuous behavioural monitoring rather than reliance on point-in-time validation.

> Control design is not a documentation exercise. It is an operational discipline.

Monitoring, Drift Detection and Behavioural Assurance

Monitoring is the operational core of AI governance in production and the capability most organisations underinvest in relative to its importance.

AI systems must be monitored across multiple dimensions simultaneously. Performance monitoring tracks whether outputs are being produced at the expected level of accuracy and reliability. Data drift monitoring detects changes in input distributions that could invalidate model predictions. Concept drift monitoring identifies situations where the relationship between inputs and correct outputs has shifted, even when distributions remain stable. Output monitoring reviews actual decisions or content produced by the system for signs of bias, safety concern or unexpected patterns.

Effective monitoring requires more than tracking isolated metrics. It requires defining a behavioural envelope for each AI system. This is the boundary within which the system is expected to operate across dimensions such as accuracy, bias thresholds, refusal rates, response quality, latency, cost and drift tolerance. Monitoring then becomes the continuous validation that the system remains within this defined envelope. When a system moves outside this envelope, it is not simply a technical deviation. It is a governance event that requires escalation, assessment and potential intervention.

Each monitoring dimension requires defined thresholds, automated alerting and clear escalation paths. A monitoring signal that generates an alert but no response is not governance. The escalation path must connect the technical signal to the person or forum with the authority to act.

Consider a retailer using an AI-driven demand forecasting system. Monitoring detects that model accuracy has declined following a change in supplier patterns. The alert reaches the operations team, who assess whether the decline falls within acceptable bounds. If it does not, the issue escalates to the model owner, who initiates revalidation. If revalidation reveals that retraining is necessary, the updated model moves through release controls before deployment. Every step is predefined, traceable and accountable. Without this structure, the same decline might go undetected until inventory errors appear in financial results, at which point the failure has already caused material harm.

> Monitoring is not technical hygiene. It is the mechanism through which governance extends from approval into ongoing operation.

Evidence and Auditability

Governance without evidence is opinion. This distinction separates organisations that are genuinely in control of their AI systems from those that believe they are.

Evidence must be structured across three interconnected layers. Inventory evidence provides a current and maintained registry of AI systems in operation, with ownership, risk classification and governance status recorded for each. Without this foundation, there is no reliable view of what is being governed. Control evidence documents that controls are implemented and operating as designed, covering validation records, testing outcomes, control assessment results and exception logs. It answers whether the governance that was designed is the governance that is actually operating. Behavioural evidence records how AI systems have performed over time, including performance trends, drift indicators, threshold breaches, escalations and incident history. It answers whether the governance that is operating is working.

These three layers together enable audit readiness. Regulators conducting reviews of high-risk AI systems will ask for this evidence. Internal audit functions assessing AI governance will need access to it. Boards seeking assurance that risk is being managed will want it presented in a form that is meaningful and reliable.

> Evidence must be continuously generated and retained. It cannot be reconstructed after the fact.

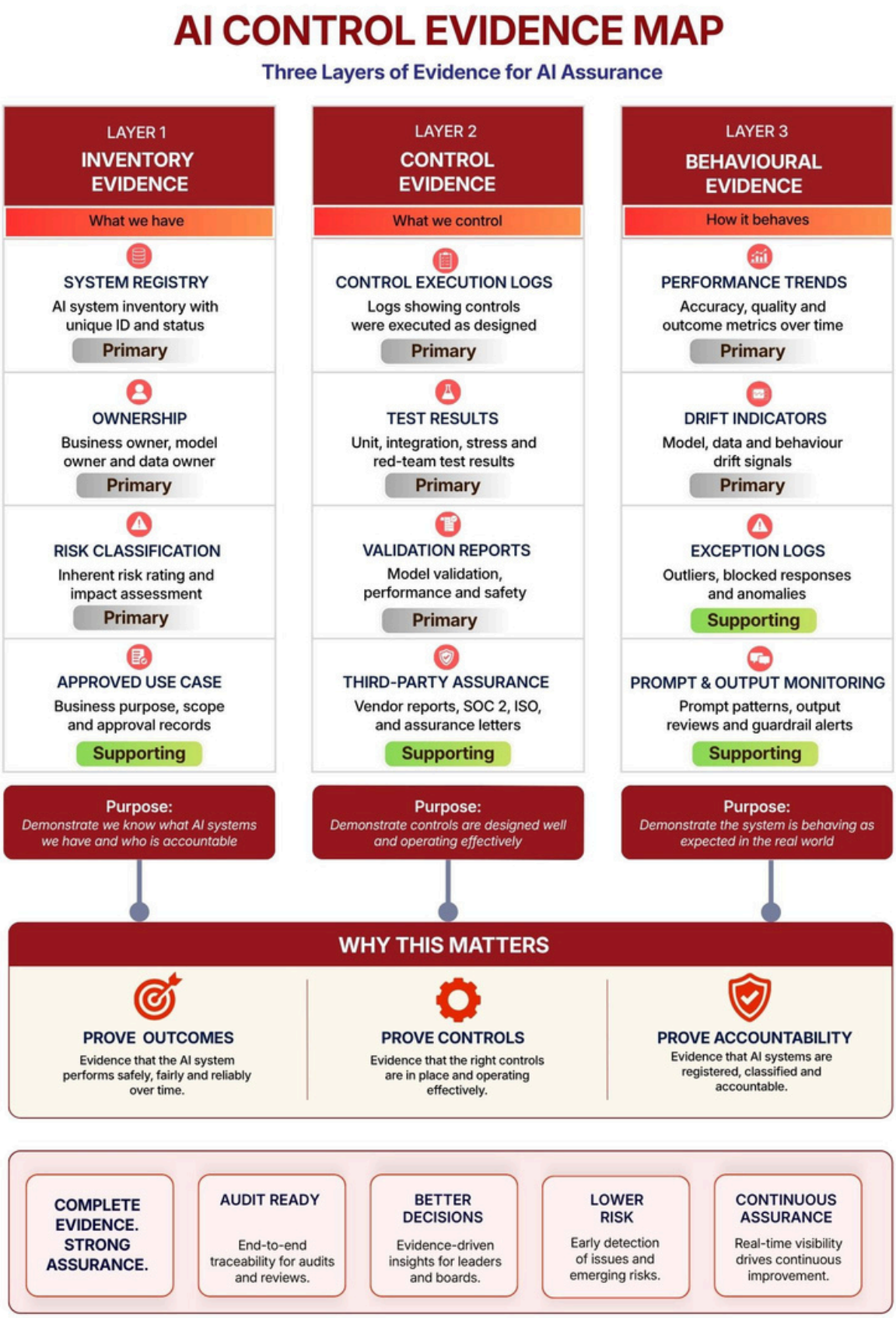


Figure 3: AI Control Evidence Map

Assurance and the Three Lines Model

The three lines of defence model provides the structural framework for AI assurance but requires meaningful adaptation to be effective in an AI context.

The first line, comprising business and technology functions, owns AI systems and is accountable for implementing and operating controls. Business owners must monitor system behaviour, maintain control documentation and escalate threshold breaches. Technology teams must build monitoring capability, maintain auditability and support the evidence requirements of the second and third lines.

The second line, comprising risk, compliance, privacy and cyber functions, is responsible for setting standards, defining risk appetite and providing oversight. Most second-line functions have not yet developed the technical fluency to assess AI-specific risks effectively. They can assess whether governance processes are being followed, but not whether systems are behaving as intended. Closing this gap requires deliberate investment in AI expertise within second-line functions.

The third line, internal audit, provides independent assurance to the board. For AI, this requires audit teams capable of assessing not only process compliance but control effectiveness and evidence quality. An audit that confirms the existence of governance frameworks does not assure the board that AI systems are under control.

The failure mode common to all three lines is treating assurance as a periodic activity. Annual reviews and point-in-time audits are insufficient for systems that operate continuously.

> AI assurance must be continuous, integrated and grounded in evidence rather than periodic, isolated and document-based.

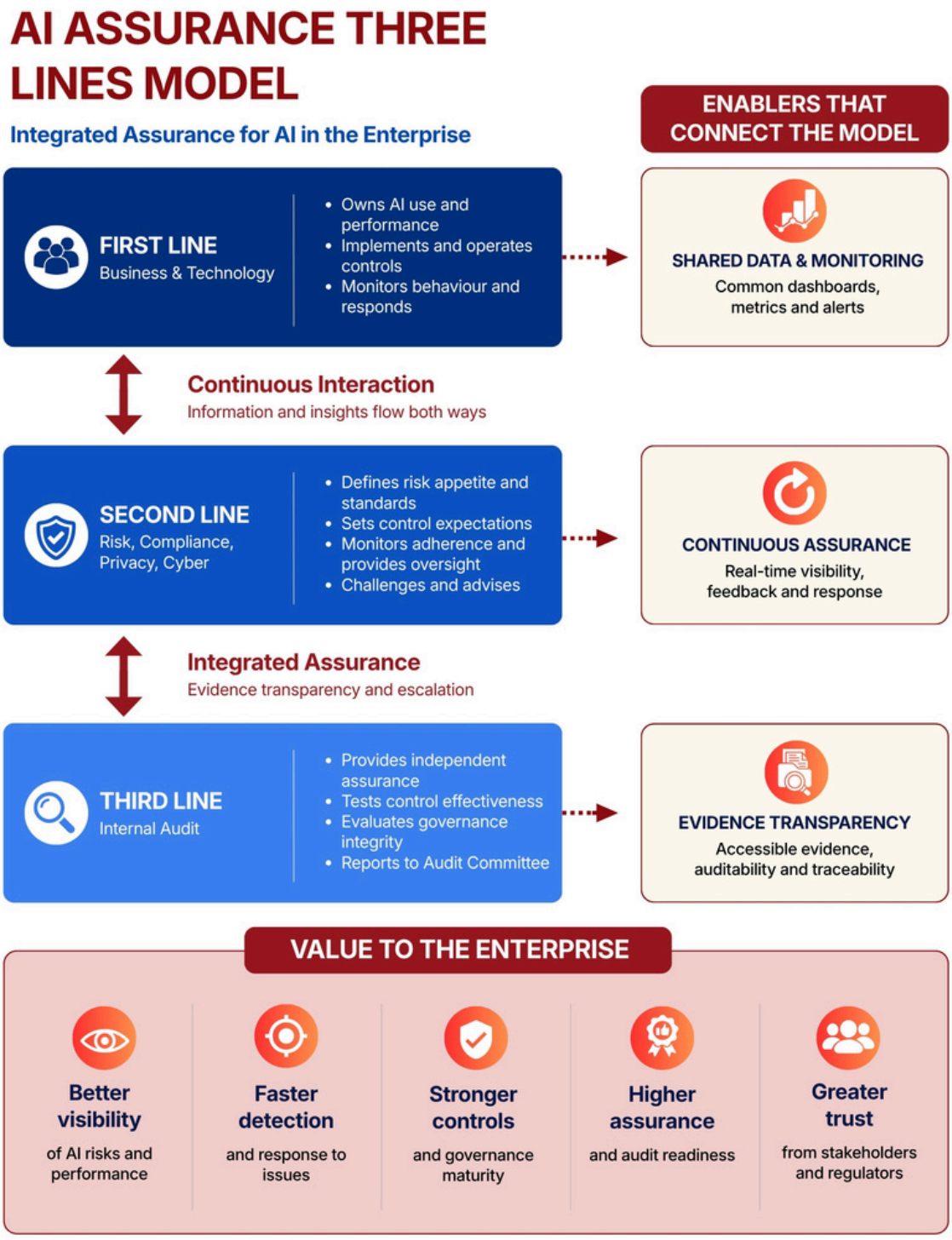


Figure 4: AI Assurance Three Lines Model

Incident Response and Remediation

AI failures will occur. What matters is how quickly they are detected, how effectively they are contained and how thoroughly the underlying causes are addressed.

Effective incident response requires predefined playbooks that establish what constitutes a reportable AI incident, who must be notified, within what timeframe and what initial containment actions should be taken. Without predefined protocols, organisations respond under pressure through improvisation, which produces slow and inconsistent outcomes.

Escalation paths must be established in advance. Root cause analysis must go beyond the immediate technical failure. Most AI incidents reflect a combination of model behaviour, data conditions, monitoring gaps and governance weaknesses. Remediation must address both the system and the control environment.

> Strong governance does not prevent all failures. It ensures failures are detected early, contained decisively and do not repeat.

Board and Executive Questions

Boards should ask:

- Can management produce a current inventory of AI systems in production, with risk classifications and governance status confirmed?
- Is monitoring in place for each system in that inventory, and is it generating regular evidence of control effectiveness?
- How are AI incidents reported to the board, within what timeframe and at what threshold of consequence?
- Does the organisation's assurance model extend into production, or does it stop at the point of approval?

Executives should ask:

- Are decision rights for AI incidents clearly defined, including escalation thresholds and response timeframes?
- Does each line of assurance have the capability to perform its role for AI, not only for conventional technology?
- Is evidence of control effectiveness being systematically generated and retained across all material AI systems?
- Could the organisation respond to a regulatory request for AI assurance evidence within a defined and reasonable timeframe?

Failure Modes

Three patterns appear consistently in organisations where AI governance has been documented but not embedded.

The Monitoring Gap

Organisations invest in model development and approval processes without building equivalent capability for production oversight. Issues accumulate without visibility. Failures surface through customer complaints, audit findings or regulatory inquiry rather than internal detection. By the time a failure becomes visible it has typically been operating for weeks or months, and the opportunity for early intervention has passed.

The Evidence Illusion

Organisations treat the existence of governance documentation as proof of control effectiveness. Documentation confirms intent. Evidence confirms reality. These are not the same thing, and regulators are increasingly clear about the distinction.

The Assurance Disconnect

Risk, compliance and audit functions operate without access to real system behaviour. Second-line functions review process compliance without access to system performance data. Third-line audit reviews governance documentation without testing whether controls are functioning. The assurance that reaches the board is structurally disconnected from the AI systems it purports to cover.

What Leaders Should Do Now

Classify risk at the point of intake, not after deployment.

Every AI use case entering the development pipeline should be assessed against a structured risk framework before work begins. This classification determines the control standard, monitoring requirements and assurance obligations that will apply for the life of the system. Retrospective risk classification, applied after deployment, is too late to shape control design.

Build production monitoring capability.

Assess honestly whether your organisation has the monitoring infrastructure to detect model drift, data shifts, output anomalies and threshold breaches in real time. For most organisations, monitoring remains the weakest link in AI governance. Investment in detection capability is not optional for organisations with material AI in production. It is the mechanism through which governance extends from approval into operation.

Define behavioural envelopes for material AI systems.

For each AI system carrying meaningful customer, financial or operational impact, establish the boundaries within which it is expected to operate, including accuracy levels, bias tolerances, latency limits, drift thresholds and escalation triggers. Without defined envelopes, monitoring has no baseline against which to measure deviation, and governance has no operational anchor.

Generate and retain evidence of control effectiveness.

Move beyond documentation of governance intent. Establish the mechanisms through which your organisation generates structured evidence that controls are operating as designed: validation records, testing outcomes, monitoring data, escalation logs and incident histories. Evidence that cannot be produced on demand does not support regulatory confidence or board assurance.

Strengthen the capability of second and third line functions.

Risk, compliance and internal audit functions cannot provide credible assurance on AI systems they lack the capability to assess. Identify where AI expertise within these functions is insufficient and address it deliberately. Second-line functions that can only assess process compliance, and audit functions that can only review documentation, will not detect control failures in production.

Define incident response protocols before incidents occur.

Establish what constitutes a reportable AI incident, who must be notified, within what timeframe and what initial containment actions are required. Organisations that improvise incident response under pressure produce slow, inconsistent and often inadequate outcomes. Predefined protocols are not administrative overhead. They are the mechanism through which control is recovered when it fails.

Closing Insight

The organisations most exposed to AI risk are not those without governance. They are those whose governance stops at approval and does not extend into operation.

AI systems do not fail at the point of design. They fail in production, where data shifts, performance degrades, monitoring is absent and accountability is unclear.

The shift required is not more governance. It is different governance.

> Governance is not a gate through which AI systems pass on their way to production. It is the system through which AI operates, continuously and at scale.

References

National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST, 2023.

National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative AI Profile. NIST, 2024.

European Union. Artificial Intelligence Act (EU) 2024/1689. 2024.

ISO/IEC 23894. Information Technology: Artificial Intelligence: Guidance on Risk Management. International Organization for Standardization, 2023.

OECD. OECD Principles on Artificial Intelligence. Updated 2024.

Board of Governors of the Federal Reserve System. Supervisory Guidance on Model Risk Management (SR 11-7). 2011.

Office of the Comptroller of the Currency. OCC Bulletin 2026-13: Model Risk Management and Emerging AI Systems. OCC, 2026.

Committee of Sponsoring Organizations of the Treadway Commission. Internal Control Considerations for Generative AI. COSO, 2026.

Institute of Internal Auditors. AI Auditing Framework. IIA, 2025.

KPMG. Deploying Trustworthy AI. KPMG, 2025.

Stanford University Human-Centered Artificial Intelligence. AI Index Report 2025. Stanford HAI, 2025.

EY. Responsible AI Pulse Survey. EY Global, 2025.

About the Author

Manoj Tavarajoo has spent over two decades working with boards, executives and senior leaders across enterprise, digital and AI transformation. His work sits at the intersection of strategy, operating model design, governance and execution, the point where good intentions either become operating reality or quietly fail.

He is the author of *Leading the AI Transformation* and *The AI Operating Model Playbook*. This paper series extends that work into the governance layer: how AI is directed, controlled, assured and held accountable at enterprise scale.

He works through MyConsultancy, an independent advisory practice based in Australia.

@manojtavarajoo

About MyConsultancy

MyConsultancy works with boards and executives navigating the distance between AI ambition and operating reality. The firm focuses on strategy, governance and operating model design, helping organisations build the portfolio discipline and transformation assurance needed to scale AI responsibly across complex enterprise environments.

www.myconsultancy.com.au

MyConsultancy