

Deloitte.

Together makes progress

Move fast—
and under control

**Engineer and purpose-build trust into
AI with automated governance**

The first in our new series on Automated AI Governance



Ask a CEO or business leader what they want from AI you'll hear a lot of answers—to **reduce cost, increase efficiency, streamline operations, protect margins, open new sources of growth**, and furthermore, to deploy and scale AI capabilities **faster with confidence**. These objectives are non-negotiable. What's less visible is the lever that makes them durable: trust. Trust in AI tools, their outputs and how it is governed. Trust is the foundation that lets an enterprise accelerate AI adoption while maintaining control and accountability. AI is truly powerful when it's proven trustworthy, auditable, and within risk tolerance.

For years, the defining question about enterprise AI was can we build it? And the answer is yes—if it's combined with a strong foundational technical architecture, data, security, functional knowledge, and AI governance and risk management framework.

But with the rise of Agentic and Frontier AI, the stakes are raised: We've built it and piloted it—but can we trust it, and can we prove it? And for boards, regulators, and customers, a policy binder or retroactive risk assessments won't be enough. Those questions will be confronted in production, at runtime, and on the day something goes wrong.



So, how can enterprises pursue that from a practical standpoint?

As organizations deploy AI at scale, whether those be traditional models, generative systems or autonomous agents, the capabilities that build trust need to be made concrete and operational as follows:

- Knowing what AI you have in your inventory, what is permitted to do against standards, who approved and who is accountable;
- Seeing how it behaves through quality, drift, and whether it is performing as intended without hallucinations and anomalous tool usage;
- Controlling what it can act on and enforcing boundaries in real time, identifying shadow AI while securing the enterprise; and
- Proving, independently, that the system proved itself, enabled by validation, tracing, and independent audit disciplines.

Underpinning all of it is the quality and integrity of the data AI depends on, the human judgment that sets the boundaries and steps in when stakes are high, and a clear organizational view of how much risk the business is willing to accept and where it must draw the line. Together, these capabilities are what allow an organization to move faster with confidence, scale what is working, and earn the trust of customers, regulators, and the board.

Therefore, trust can't just be a value to claim, but something actively engineered.

Because when AI acts on its own—deciding, transacting, calling other systems, dispatching other agents—trust can't be assumed; it needs to be designed and built in from the start, measured continuously, and enforced every time a model runs.

Does that sound like a handbrake or an accelerator? We would suggest it's the latter, because trust that's built in from the jump means you can scale AI responsibly and faster. A trust-based architecture and technology-enabled model transforms governance from a gate into enabling infrastructure with enforceable controls that help enterprises scale AI while protecting revenue, brand, reputation, and the license to operate.

Treated this way, trust becomes a strategic business enabler. And the organizations that get it right will be more effective, move faster and realize value sooner than those that treat governance as paperwork.

Glossary

AI Trust Architecture. A framework of four interlocking components: AI GRC, AI Observability, AI SPM, AI Control Plane, —built on trusted data, adequate end-to-end testing and independent audit that makes AI trust operational, not aspirational.

AI Governance, Risk and Compliance (AI GRC).

Sets policy, risk tiering, and inventory; answering who approved a system, against what standard, and whether it's still compliant.

AI Observability (AI Obs). Continuous monitoring of AI quality, drift, cost, and behavior; the difference between “we think it's fine” and “we can show it's fine.”

AI Security Posture Management (AI SPM). Discovers AI across the estate, including shadow AI, and surfaces the assets, identities, and exposures that need protecting.

AI Control Plane (AI CP). The runtime layer where policy becomes action; guardrails, permissions, and kill switches enforced the moment an AI system runs.

First, second, and third lines. The three-lines-of-defense model applied to AI: the team that builds and traces the system (first line), the risk function that independently challenges it (second line), and internal audit that verifies it to the board (third line).

Frontier AI. The most advanced, capable AI models operating with greater autonomy and risk than traditional models that have been widely deployed and tested over time.

Guardrails. The deterministic, rules-based controls, written down in advance that bound what an AI system is permitted to do.

Human-in-the-Loop (HITL): A person must approve each AI-proposed action before it executes, serving as gatekeeper for high-stakes, hard-to-reverse decisions.

Human-on-the-Loop (HOTL): The AI system acts autonomously while a person supervises via dashboards/alerts and intervenes only on exceptions, suited to high-volume, time-sensitive use cases.

Human-in-Command (HIC): Humans retain ultimate authority over whether, how, and within what boundaries an AI system is deployed, independent of any single decision.

Trustworthy AI™. Deloitte's proprietary framework of principles for responsible AI, which the AI Trust Architecture puts into operational practice.

Why now? The trust inflection point has arrived.

Four forces have turned AI trust from an aspiration into an operational imperative.

- **Evolving regulations, guidance, and laws.** The EU AI Act, global regulators, and standards like the NIST AI RMF and ISO/IEC 42001 now expect demonstrable, ongoing governance—and not a one-time sign-off.
- **AI inherits traditional technology risk and adds new failure modes.** AI remains exposed to software, cloud, identity, and supply chain weaknesses while adding probabilistic behavior, drift, and outputs that are confidently wrong. Agentic systems compound this by planning, chaining tools, and acting across systems faster than any human can review.
- **Sprawl outruns oversight.** Models, vendors, copilots, and shadow AI—now joined by citizen-built agents—multiply faster than manual review can track. A human in the loop will not be sufficient.
- **AI expands the attack surface.** Models, agents, tools, data stores, connectors, and machine identities create new paths for prompt injection, credential misuse, data exfiltration, poisoning, and supply chain compromise.

And a fifth force compounds the first four—cost. As AI moves from pilots to production, spend on compute, tokens, agents, and the specialized talent to run them is rising so fast that it's on the CFO's agenda, not just that of the CIO. Left ungoverned, that growth can quietly turn an efficiency initiative into a cost center of its own.

Solution? A deliberate, pragmatic approach to embedding trust through the AI Trust Architecture keeps growth and discipline aligned with reusable controls that avoid costly reworks. As a result, spend tends to be routed toward the use cases that have earned it. (Read more in Deloitte's "Trustworthy AI™ Cost" point of view, coming soon.)

Geographic rulebooks are changing as fast as AI itself.

In the EU, the Digital Omnibus¹ has deferred the AI Act's high-risk obligations—stand-alone (Annex III) systems to December 2027 and embedded systems to 2028—even as its transparency duties land on schedule. **In the US**, this is still no comprehensive federal statute and a widening patchwork of state laws. NIST's AI Risk Management Framework keeps maturing through new profiles (both Generative AI and cyber AI). ISO/IEC 42001 anchors certifiable governance, and **the map is global**: the UK's principles-based stance, China's Generative AI and algorithm rules, South Korea's AI Basic Act, and fast-moving regimes across Canada, Brazil, Japan, and Singapore.

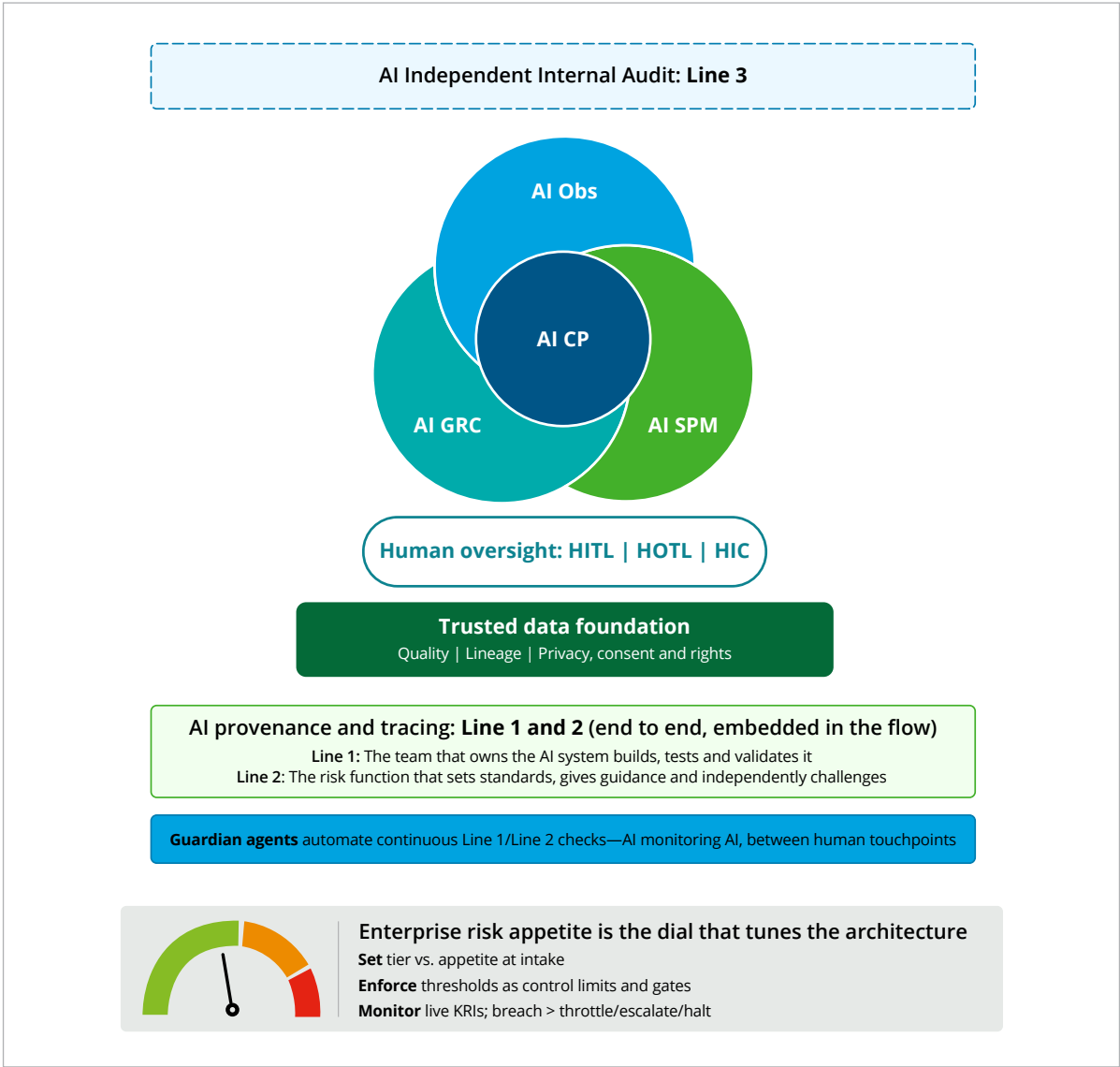
Sector regulators are picking up speed. For example, in financial services, the US Treasury's Financial Services AI Risk Management Framework and AI Lexicon (2026) and the Financial Stability Board's consultation on 12 "sound practices" for responsible AI adoption (2026) are both anchored in board-level risk appetite, life cycle governance, and human oversight. Therefore, building a unified, mappable control set will be the most pragmatic way to approach.

The *AI Trust Architecture*— living and active guardrails

Responsible AI principles tell you what good looks like. They rarely tell you where it lives. The AI Trust Architecture does: It lives in the systems, workflows, controls, evidence, and accountable human decision points. To operationalize trust, organizations need capabilities that define governance requirements, enforce them at runtime, continuously monitor outcomes, and independently verify performance. Collectively, these capabilities form the AI Trust Architecture.

The AI Trust Architecture proposes four interlocking components, each answering a different question about an AI system.² This interdependence rests on trusted data; is traced and tested end to end; and is bound by independent internal audit, with people accountable throughout (figure 1). Governance, cybersecurity, and resilience span every component. Let’s unpack it.

Figure 1. The AI Trust Architecture: Four components on a trusted data foundation, with embedded AI provenance and tracing (lines 1 and 2), independent internal audit (line 3), and human oversight throughout.



These components represent enterprise capabilities and operating functions that may be enabled by one or more technologies. The value comes from how they work together as an integrated system, not from any single platform.

- **AI Governance, Risk and Compliance (GRC).**

Sets policy and risk tiering, and maintains an inventory that connects each use case to its models, agents, owners, identities, tools, data sources, connectors, permissions, dependencies, environments, and third parties. It provides defensible answers to these questions: Who approved the system? Against what standard? Has it remained compliant?

- **AI Observability (AI Obs).**

Continuous telemetry on quality, drift, hallucination, cost, security behavior, anomalous tool use, privilege escalation, policy bypass, and data exfiltration, with relevant signals integrated into enterprise monitoring capabilities. Said another way: The difference between “we think it’s fine” and “we can show it’s fine.”

- **AI Security Posture Management (AI SPM).**

Discovers AI across the estate, including shadow AI, and identifies assets, identities, data flows, dependencies, misconfigurations, and exposures. AI SPM informs prioritization, while runtime protection, threat detection, response, and recovery are delivered through the AI Control Plane and security operations.

- **AI Control Plane (AI CP).**

The runtime fabric where policy becomes action through guardrails, agent identity and authorization, least-privilege and tool-level permissions, short-lived credentials, routing, transaction limits, sandboxing, containment, and a kill switch. Where intent finally touches reality.

- **Trusted Data Foundation.**

Governs quality, lineage, privacy, consent, and rights. Trust depends on the data beneath it; if the inputs are unreliable, everything built on top starts to feel like a guess.

- **AI Provenance and Tracing (encompassing model risk management and other risk functions where applicable).**

End-to-end testing by the team that owns the system (first line) and independent challenge by the risk function (second line), before go-live and continuously after, potentially with guardian agents automating these checks at agentic speed.

- **Human oversight: The oversight of an accountable core.**

Human in the loop (HITL) to approve high-stakes, hard-to-reverse actions; human on the loop (HOTL) to supervise autonomy and step in on exceptions; human in command (HIC) to set boundaries and own the outcome. The riskier and less reversible the action, the closer the human needs to be.

- **Independent internal audit.**

Independent review that pulls evidence from every component to prove, to a board or a regulator, that the system is operating within defined risk tolerance.

Governance without enforcement is a memo. Enforcement without visibility is a guess. And visibility without security is exposure.

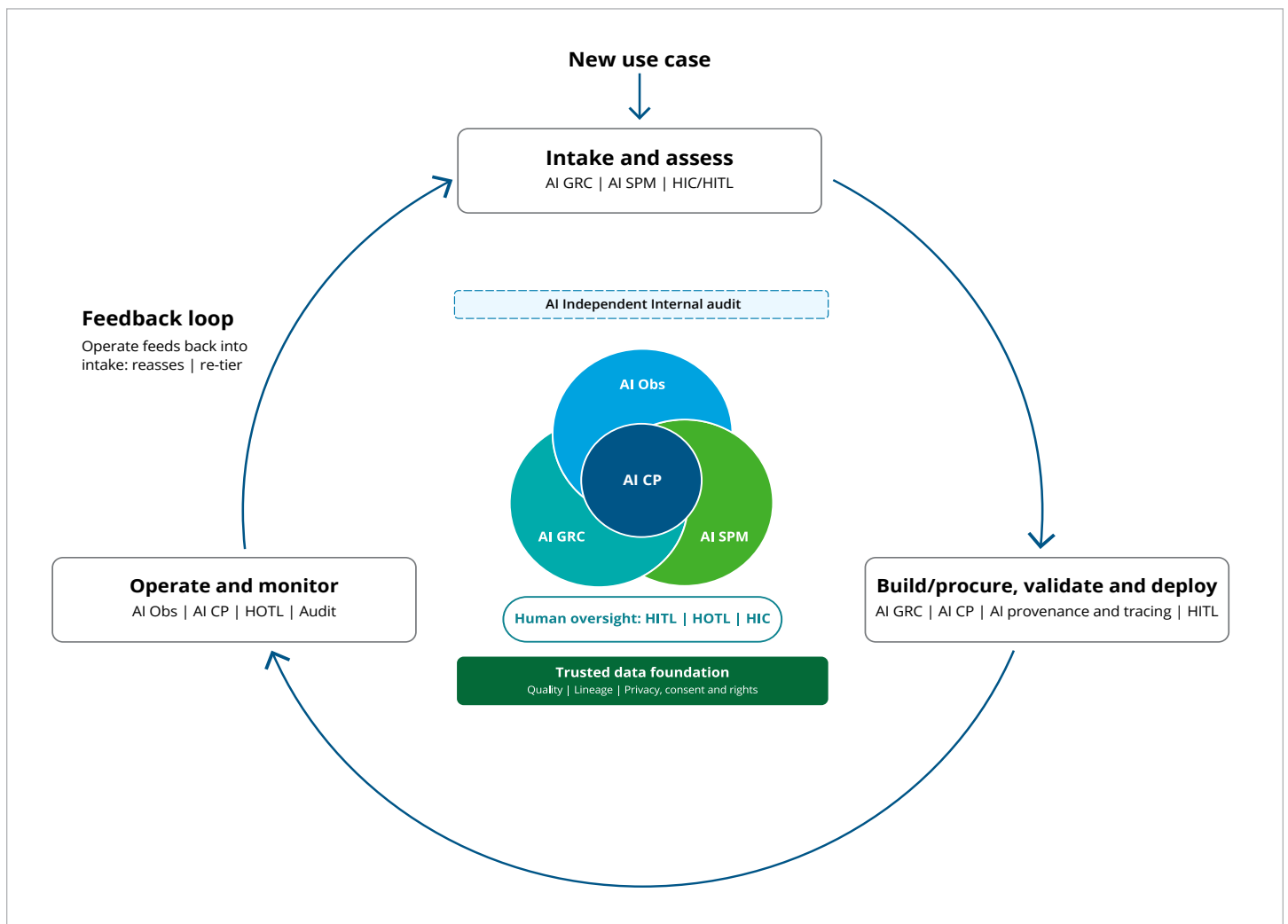
We hope the power of putting theory into action is starting to make sense: AI GRC defines, the AI Control Plane enforces, AI Observability verifies, and AI SPM discovers and protects. Insights from AI Observability and AI SPM—including runtime protection, detection, response, and recovery—flow back into policy and runtime controls. Remove any one component and the loop breaks.

Trust is not a gate you clear at launch. It is a loop you run for the life of the system—across intake, build, and operate—with what you learn in production feeding straight back into how you assess and control what comes next (figure 2). And it doesn’t end there: secure-by-design engineering, adversarial testing, third-party and supply chain assurance, containment, rollback, recovery, and forensic readiness must run through that life cycle.

Not four tools— *one continuous loop*

*And the weakest link is the limit
on how far you can trust and scale*

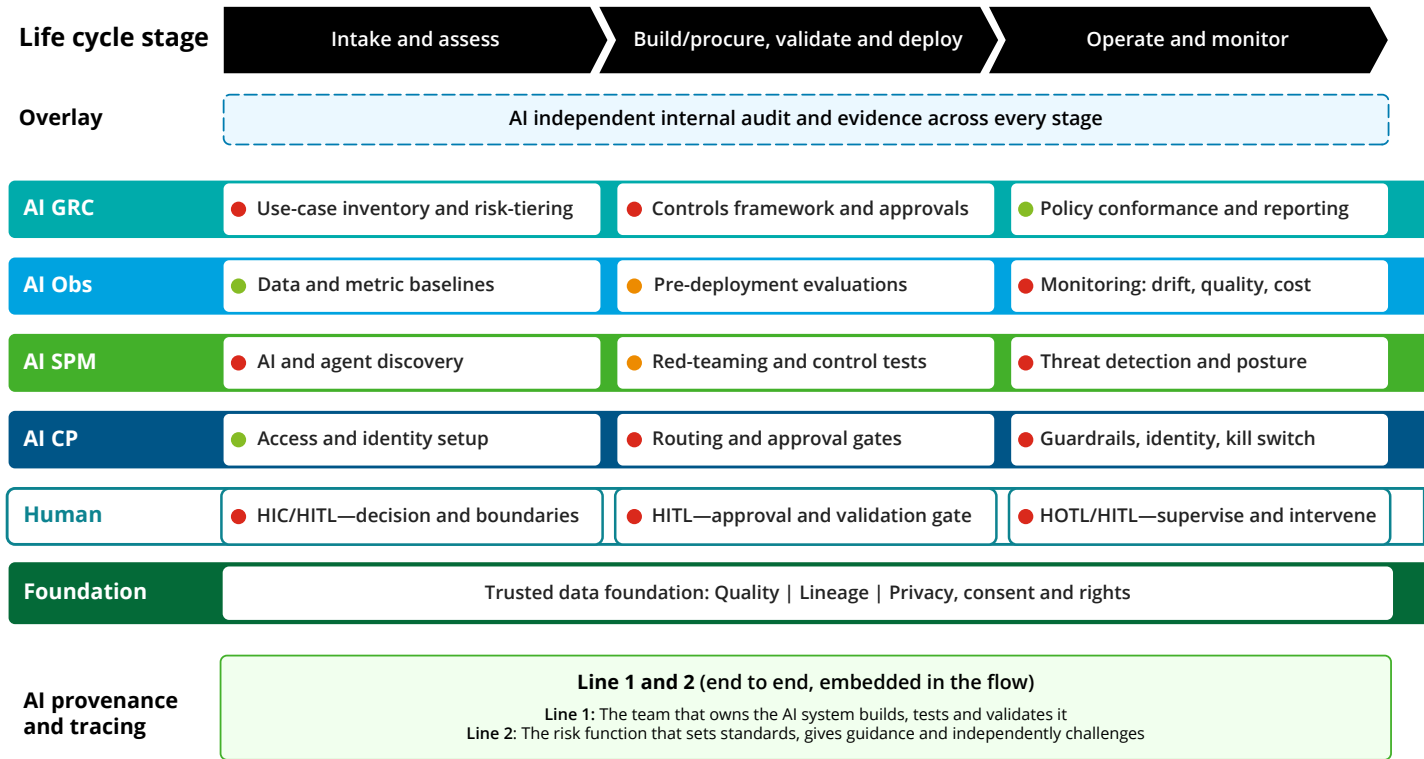
Figure 2. The architecture runs continuously across the AI life cycle—operate and monitor feeds back into intake and assess.



Zoom in to any turn of that loop and the same four components are at work—leading in some stages, supporting in others, with human oversight throughout (figure 3).

AI provenance and tracing runs end to end. The team that owns each AI system builds, tests, and traces it (the **first line**); the risk function sets standards and guides and independently challenges (the **second line**); and independent internal audit provides a periodic, independent sidecheck (the **third line**).

Figure 3. Component activity across the AI life cycle—the color red marks the primary component at each stage.



Intensity of each component's activity at that stage: ● Primary ● Supporting ● Light

Turning the *dial of risk* up and down

Every control in this architecture is ultimately decided by one question: **How much risk are we willing to assume?** And only your senior leadership and board can answer that.

What is their appetite for risk? Answering that turns oblique questions such as “Is this safe?” into something measurable. It establishes thresholds about where an AI system may run freely, where it needs a human in the loop, and where it must stop. A graduated autonomy framework translates those thresholds into operating boundaries—adjusting permitted actions and human oversight based on privilege, data sensitivity, system criticality, reversibility, external impact, and potential blast radius.

As illustrated in Figure 3, these thresholds **enter at the intake and assess stage**, when a use case is risk-tiered against the enterprise’s risk appetite; they **harden into control limits** as the system is built and routed through the AI Control Plane; and they become **live tripwires in production**, where AI Observability watches drift, error, exposure, and cost and steps in—and the tripwire tells the system to slow, escalate, or halt.

Set your risk too narrow and AI never leaves the lab; set it too wide and you’re one silent failure from the headlines. The answer? Treat risk appetite as a living threshold, built into how AI is tiered, deployed, and monitored.

Governing at the *speed of AI*—with AI

Guardrails cover the rules you can write down. But as autonomous agents multiply across the enterprise, something must watch the agents themselves. Drawing on Gartner's classification³ and Deloitte's implementation experience, three kinds of guardian agents operate across the enterprise AI stack:

- **Monitors** observe and track AI and agentic actions—overseeing workflows, detecting anomalies, and flagging issues for human or AI-based follow-up in real time.
- **Reviewers** identify and review AI-generated output and content for accuracy and acceptable use—validating summaries, outputs, and decisions before they reach the next stage.
- **Protectors** adjust or block AI and agentic actions using automated intervention during operations—responding promptly to rogue agent activity, policy breaches, or threshold violations.

All three of these guardians work in concert with guardrails, the live outputs of the four components, and the enterprise's risk thresholds. Depending on the complexity and risk of a workflow, guardian agents may be generic and enterprise-wide—spanning the estate to enforce policy and check drift—or purpose-built for a single high-stakes use case.⁴ Because guardian agents may hold elevated privileges, they must be governed as high-risk systems, with separate identities, least-privilege access, deterministic boundaries, segregation of duties, tamper-resistant logging, independent validation, and clear human escalation. Guardian agents supplement rather than replace deterministic controls.

Grounded in *Trustworthy AI*[™]

The AI Trust Architecture we've been laid out operationalizes Deloitte's Trustworthy AI[™] principles and gives the major frameworks a home—and a confidence throughout the contexts in which your enterprise operates. For example, **AI GRC** maps to the NIST AI RMF Govern and Map functions and ISO/IEC 42001; **AI Observability** to Measure and Manage; the **AI Control Plane** to runtime enforcement and the EU AI Act's human-oversight duties (Article 14); and cybersecurity and resilience (**AI SPM**) span all four components and the trusted data foundation, consistent with the direction of NIST's initial preliminary draft Cyber AI Profile.

Where to *start*

In the midst of the pressure to do something, many organizations are installing AI on top of current architectures (systems, technology, processes, data) to achieve what they believe will be faster results. But this boxes them into legacy processes, leaving them with a limited runway for reimagining the ways to do work. Seeking a shift in AI-native thinking and architecture—including a drastically different way to build next-generation enterprises—will help to take the vision farther than currently capable.

Other enterprises already own fragments of what we've been discussing, but not a comprehensive approach; a model inventory here, a monitoring tool there, some guardrails in one application. The challenge is to architect deliberately for the agentic sprawl across the enterprise ecosystem and connect those fragments into one system for trust. Clear cross-cutting of functional roles and responsibilities are essential. Without defined ownership across business lines, risk functions, and technology teams, the governance is likely to fragment again.

- **Define first.** Stand up one risk-tiered inventory connecting AI use cases to models, agents, owners, identities, tools, data sources, connectors, permissions, dependencies, environments and third parties, creating the reference point for everything else to align to.
- **Enforce by default.** Route generative and agentic use cases through a control plane so guardrails, identity, authorization, least-privilege access and tool-level permissions apply consistently.
- **Observe early.** Instrument high-risk use cases for quality, telemetry, thresholds, alerts, and run-time evidence that are available from day one.
- **Protect continuously.** Discover shadow AI, harden against AI-specific threats, manage third-party and supply chain exposure, and prepare for containment including credential revocation, rollback recovery, and forensic investigation. Wire the findings back into policy and runtime controls.

Regulation will keep moving; the AI Trust Architecture is how you stop rebuilding for each new rule and start mapping one control set to whatever arrives. Do this well, and **trust becomes the strategic enabler to scale AI.**

How Deloitte can help you *automate* your AI governance

Recognized by the International Data Corporation (IDC) as the worldwide No. 1 AI services provider for the fourth consecutive time (2025 IDC MarketScape⁵), Deloitte brings the full depth of that capability to the AI Trust Architecture. From strategy through daily operations, we can help organizations close the gap between claiming governance and running it:

- **AI governance operating model and policies.** Design and implement the risk management program; governance frameworks inclusive of cross-cutting stakeholders with clear roles and responsibilities; and operating model, controls, and board-level reporting that turn AI principles into enforced, auditable practice.
- **Agentic AI strategy and life cycle process design.** Design the strategy and operating model for risk-tiered use cases, including graduated autonomy, governance processes, and human-oversight modes that let agents operate safely at scale.
- **Tech selection and implementation.** Navigate a crowded, fast-consolidating vendor market to select fit-for-purpose capabilities per component, including build-versus-buy considerations, and implement them with testing, validation, and compliance controls.
- **Guardian agents.** Design, build, and implement guardian agents tailored to your highest-risk agentic workflows, automating the continuous checks the architecture requires at agentic speed. Govern the guardian agents themselves through separate identities, least-privilege access, deterministic boundaries, logging, validation, and human escalation.
- **AI provenance and tracing.** Enhance reliability through end-to-end testing, validation, and continuous performance monitoring.
- **Cyber and AI security.** Secure AI systems across the life cycle through architecture and engineering, agent and non-human identity, adversarial testing, runtime protection, threat detection, third-party and supply chain risk, incident response, resilience, and SOC integration.
- **Organizational change, training, and communications.** Build AI literacy and governance culture to bridge AI technical skill gaps with governance tool implementations.

The organizations that succeed won't choose between moving fast and staying in control. They'll do both—and prove it on demand. That is what a technology-enabled governance model for trust makes possible.

Authors

Gowri Zoolagud

Managing Director
AI Governance, Risk Management
and Technology Enablement Leader
Deloitte & Touche LLP
gzoolagud@deloitte.com

Derek Snidauf

Principal
AI Risk Management and Governance Leader
Deloitte Transactions and Business Analytics LLP
dsnidauf@deloitte.com

Clifford Goss, PhD

Partner
AI Trust Leader
Deloitte & Touche LLP
cgoss@deloitte.com

Prakul Sharma

Principal
AI Data and Engineering Leader
Deloitte Consulting LLP
praksharma@deloitte.com

Mehdi Houdaigui

Principal
AI Cybersecurity Leader
Deloitte & Touche LLP
mhoudaigui@deloitte.com

Toan Viradet

Managing Director
AI Human + Agent Operating Model Leader
Deloitte Consulting LLP
xviradet@deloitte.com

Our Industry Champions

FSI: Brian Merrill, Managing Director, Deloitte Transactions and Business Analytics LLP
bmerrill@deloitte.com

LSHC: Michael Crowthers, Managing Director, Deloitte & Touche LLP
mcrowthers@deloitte.com

TMT: Richard Tumber, Principal, Deloitte & Touche LLP
ritumber@deloitte.com

ER&I: Howard Friedman, Managing Director, Deloitte & Touche LLP
hfriedman@deloitte.com

Consumer: Derek Snidauf, Principal, Deloitte Transactions and Business Analytics LLP
dsnidauf@deloitte.com

GPS: Joseph Conti, Managing Director, Deloitte & Touche LLP
joconti@deloitte.com

1. Regulation-AI.eu, "[EU AI Act Omnibus 2026: New compliance deadlines explained](#)," June 18, 2026.
2. AI system is broadly termed inclusive of technology (legacy and new), data, processes, and people that includes different archetypes of AI in varying capacities.
3. Gartner, "[Market guide for guardian agents](#)," February 25, 2026.
4. Read more about Deloitte's guardian agents in our new paper, "[Taking a first step: Guardian agents for agentic AI applications](#)."
5. "[Deloitte, IDC MarketScape: Worldwide Artificial Intelligence Services 2025 Vendor Assessment](#)," August 19, 2025.



About Deloitte

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee, and its network of member firms, each of which is a legally separate and independent entity. Please see www.deloitte.com/about for a detailed description of the legal structure of Deloitte Touche Tohmatsu Limited and its member firms. Please see www.deloitte.com/us/about for a detailed description of the legal structure of Deloitte LLP and its subsidiaries. Certain services may not be available to attest clients under the rules and regulations of public accounting.

This publication contains general information only and Deloitte is not, by means of this publication, rendering accounting, business, financial, investment, legal, tax, or other professional advice or services. This publication is not a substitute for such professional advice or services, nor should it be used as a basis for any decision or action that may affect your business. Before making any decision or taking any action that may affect your business, you should consult a qualified professional advisor. Deloitte shall not be responsible for any loss sustained by any person who relies on this publication.