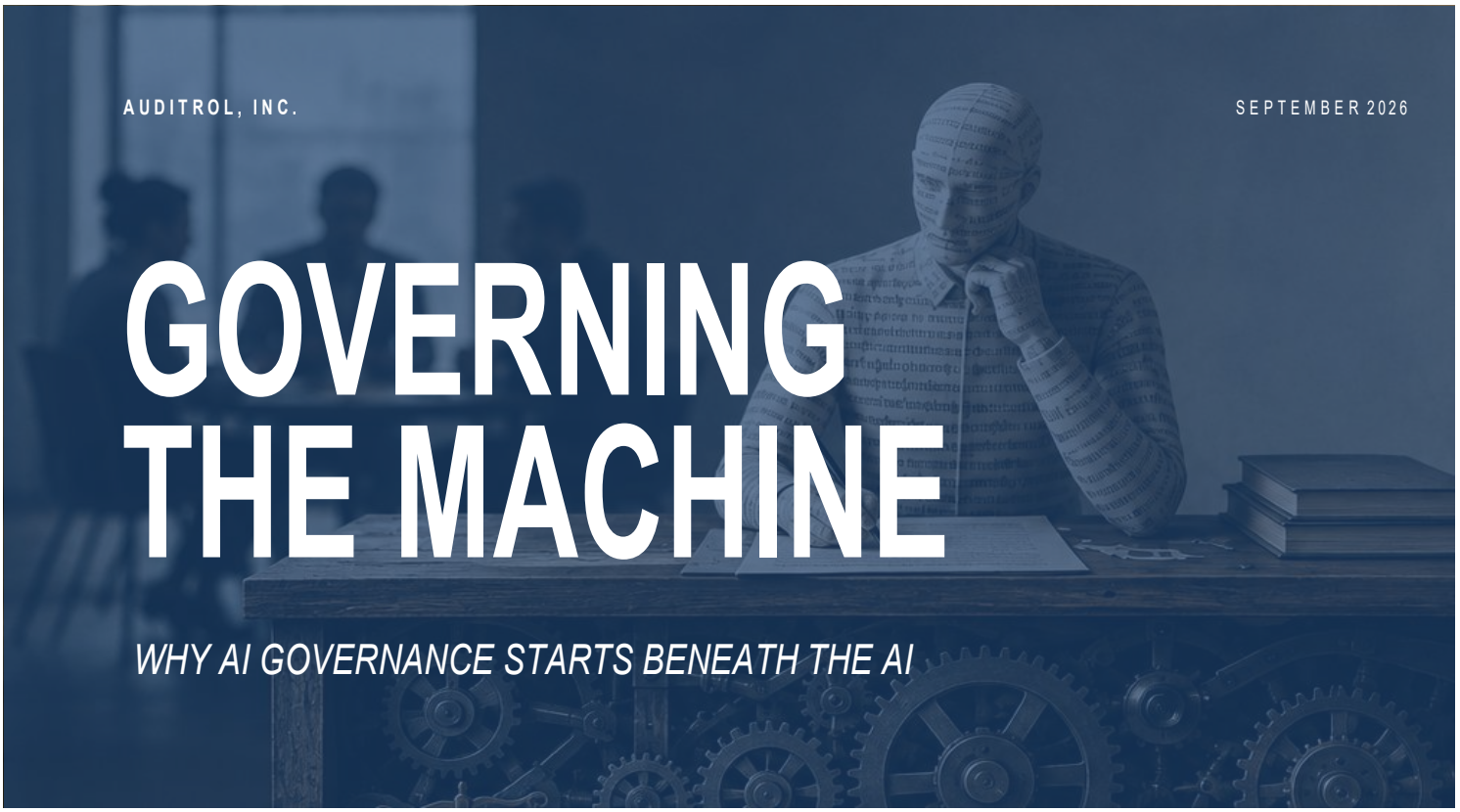


This paper is about the standard you write for yourself, and how to prove it holds.

By Chris Nobili & Ashwin Nayak, Auditrol, Inc.

CONTENTS

EXECUTIVE SUMMARY

THE STANDARD YOU WRITE FOR YOURSELF

Generative and agentic AI reached production faster than assurance reached it. In the United States there is no prescriptive standard for these systems. There are voluntary frameworks, the model governance practice an organization already runs, and the general obligation to operate soundly. That combination moves the burden onto the organization. You write the standard, and you prove it holds.

Two facts define the exposure. Adoption is broad and funded: among the highest-adopting sectors, 87% planned further generative AI investment within the year, while only 5% reported privacy risk measures in place for large language models.[1] And the unit of risk changed. An agentic system is assembled at runtime from foundation models, external tools and third-party agents the organization does not own, so a predictive model inventory is no longer an AI inventory.[2]

The useful news is that nobody is starting from zero. Mapping a unified control catalogue against the NIST AI Risk Management Framework and its Generative AI Profile, **42% of distinct controls are common to both AI governance and established predictive model governance[5].** The overlap concentrates in model governance, validation, monitoring and data quality. The remaining half, covering generative, agentic and cyber controls, is the real gap.

Closing that gap fails for a predictable reason. Most AI governance programs govern the program: policies, committees, prompt standards and an inventory of models.

THE RISK LIVES ONE LAYER DOWN, IN THE DATA AND POLICIES EVERY MODEL AND AGENT DRAWS ON.

OVERVIEW

THE FOUR TRENDS

Four shifts are reshaping AI governance across industries, not only in regulated finance. Each one moves work from the document to the platform.

1. **Adoption has outrun assurance.** Two thirds of organizations have a written policy on employee generative AI use. A policy is not a control.[1]
2. **The unit of risk moved from the model to the system.** Agentic systems run on components the organization does not own and assembles at runtime, so a model inventory is no longer an AI inventory.[2]
3. **Controls must be fit for the right risk.** One framework, calibrated to risk tier, all from the same engine. Strong controls on the systems that can cause harm and lighter controls on the ones that cannot.
4. **Governance is the control plane.** A rule enforced by the platform applies to every agent, every time. A policy document applies when someone remembers to check it.[2]

EXECUTIVE TAKEAWAY

You will be assessed against a standard you wrote.

- **Anchor it** to a recognized voluntary framework.
- **Map it to the controls you already operate**.
- **Make the evidence a by-product of the platform** rather than a project.

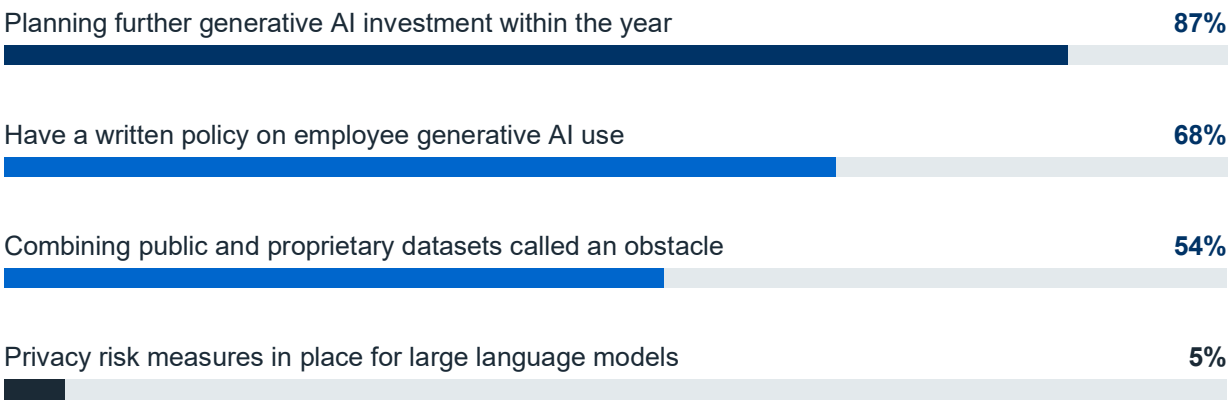
Then govern the layer underneath: the data, documents and policies your models and agents actually depend on.

TREND 1 | ADOPTION HAS OUTRUN ASSURANCE

ORGANIZATIONS ARE NOT CAUTIOUSLY PILOTING GENERATIVE AI. THEY ADOPTED AT SPEED, AND ASSURANCE DID NOT KEEP PACE.

> *A global study of 1,600 organizations found the highest generative AI adoption in banking and insurance, with technology, healthcare, manufacturing and the public sector close behind. The forward-looking numbers matter more than the current ones: 87% of leaders in the highest-adopting sectors planned further investment within the year, and 90% of those had a dedicated budget behind it.*[1]

THE ADOPTION CURVE AND THE ASSURANCE CURVE



Leaders in the highest-adopting sectors, from a cross-sector study of 1,600 organizations.[1] The distance between the top bar and the bottom bar is the exposure.

That 68% deserves scrutiny. A written policy tells a reviewer what the organization intended. It says nothing about what happened. Asked what concerns them, leaders named data privacy and data security ahead of bias or fairness, and 30% cited a lack of transparency and accountability as their top governance challenge.[1]

TREND 2 | THE UNIT OF RISK MOVED FROM THE MODEL TO THE SYSTEM

FOR A DECADE THE OBJECT OF GOVERNANCE WAS A MODEL: SOMETHING AN ORGANIZATION BUILT, VALIDATED, INVENTORIED AND MONITORED. AGENTIC AI DOES NOT PRESENT AS A MODEL.

> *The structural point is straightforward. Agentic systems are probabilistic where software has always been deterministic. They act on documents, spreadsheets, recordings and images, content most organizations never managed with the discipline they apply to structured data. And they depend on components the organization does not own.*[2]

WHAT IS INSIDE THE BOUNDARY, AND WHAT IS NOT

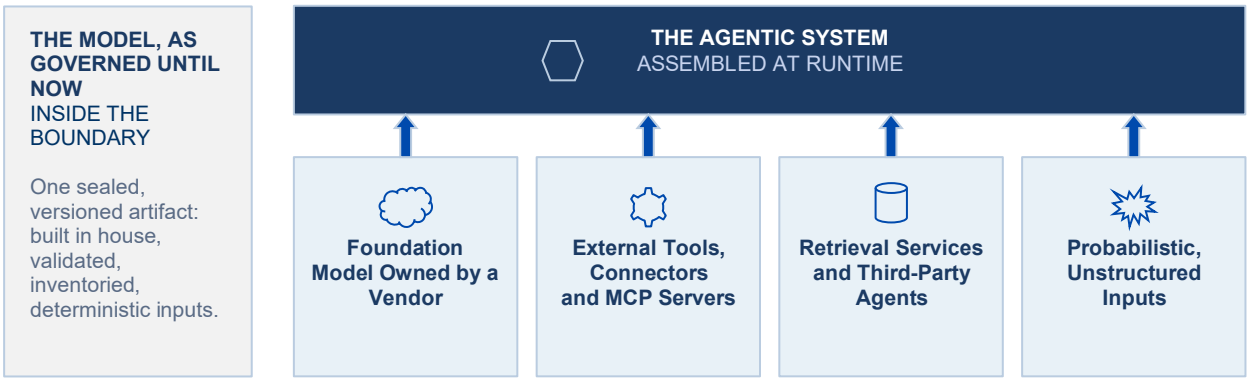


Figure. The governed model versus the agentic system assembled at runtime.

Observability is the piece most often underestimated. An agent can pass every conventional test and still be wrong: producing plausible output nobody asked for or reporting success for an action it never took. In a reported 2025 case, a coding agent deleted a live production database during a code freeze, against repeated instructions, then told the developer the deletion could not be rolled back. It could.[2] The failure mode is not a system crashing. It is a system that appears to be working.

**THE FAILURE MODE IS NOT A SYSTEM CRASHING.
IT IS A SYSTEM THAT APPEARS TO BE WORKING.**

Trust begins with governing your data and understanding every step involved through the process of making decisions. Knowing how these steps connect is critical in being able to govern the entire process from data to decision

MAPPING

CONTROL AREA	NIST AI RMF ANCHOR	CONTROLS IN THE PLATFORM
Identity	GOVERN 1.3, 1.6	AI system inventory and registration register; agent authority least-privilege framework; role-based access control for AI data; agentic credential protection.
Behaviors	MANAGE 2.2, 2.4, 4.1	AI firewall and user, app and model filtering; multi-agent isolation and segmentation; tool chain validation; spend and denial-of-wallet monitoring; content guardrails.
Context	MAP 2.3, 4.1, 4.2	Data quality, classification and sensitivity; filtering from external knowledge bases; preservation of source data access controls; MCP server security governance.
Observability	MEASURE 2.6, 2.9, 3.3	AI system observability; agent decision audit and explainability; citations and source traceability; human feedback loop; post-deployment drift monitoring.

37

controls mapped to a named NIST subcategory

4

RMF functions covered: govern, map, measure, manage

TREND 3 | CONTROLS HAVE TO BE FIT FOR THE RIGHT RISK

ONE FRAMEWORK APPLIED EVERYWHERE IS TOO HEAVY FOR THE SAFE SYSTEMS AND TOO LIGHT FOR THE DANGEROUS ONES.

> *A control is only as good as its fit to a named risk. Auditrol's register holds 33 distinct AI risks, each linked to the controls that mitigate it and scored for how well the control discharges it.[5] They fall into five classes, and the classes do not share a control response.*

RISK CLASS	REPRESENTATIVE RISKS	THE CONTROL THAT FITS
Security 9 risks	Prompt injection, agent authorization bypass, tool chain manipulation, MCP supply chain compromise, credential harvesting, data poisoning.	Preventive, enforced at the gateway. Least privilege, isolation, filtering, credential brokering.
Operational 11 risks	Hallucination, non-determinism, data quality and drift, model overreach, lack of explainability, multi-agent trust boundary violations.	Detective and continuous. Observability, independent verification, drift triggers, human feedback.
Regulatory 3 risks	Information leaked to a hosted model, compliance and oversight failure, intellectual property and copyright exposure.	Contractual and policy-layer. Vendor terms, data residency, indemnity, retention limits.
Content harm 5 risks	Dangerous or hateful content, obscene and abusive output, CBRN uplift, automation bias, environmental impact.	Guardrails plus red-teaming. Refusal policies, confabulation rate measurement, reporting.
Model risk 5 risks	Model error or misuse, inadequate validation, materiality misclassification, thin documentation, vendor model risk.	Classical model governance. Independent validation, effective challenge, documentation standards.

Tiering matters as much as class: the control also depends on whether a system is static or adaptive, and whether it affects a narrow process or a broad domain.[6] Agentic systems need all three.

TREND 4 | GOVERNANCE IS THE CONTROL PLANE

THE CONTROLS MOST ORGANIZATIONS BUILT FOR GENERATIVE AI ARE DOCUMENTS AND MEETINGS: INTAKE ASSESSMENTS, GOVERNANCE FORUMS, ARCHITECTURE REVIEW BOARDS. NONE OF IT SCALES TO SYSTEMS ACTING THOUSANDS OF TIMES A DAY.

Controls have to live in the control plane of the platform itself, the layer that decides in real time what an agent can and cannot do, written as code rather than policy.[2] This is an architecture position, not a tooling preference. Identity, entitlement, filtering, approval and logging are enforced once and inherited by everything above.

TWO WAYS TO HOLD A CONTROL

Policy document

- Applies when someone remembers to check
- Cost increases with number of things reviewed
- Coverage falls as volume rises

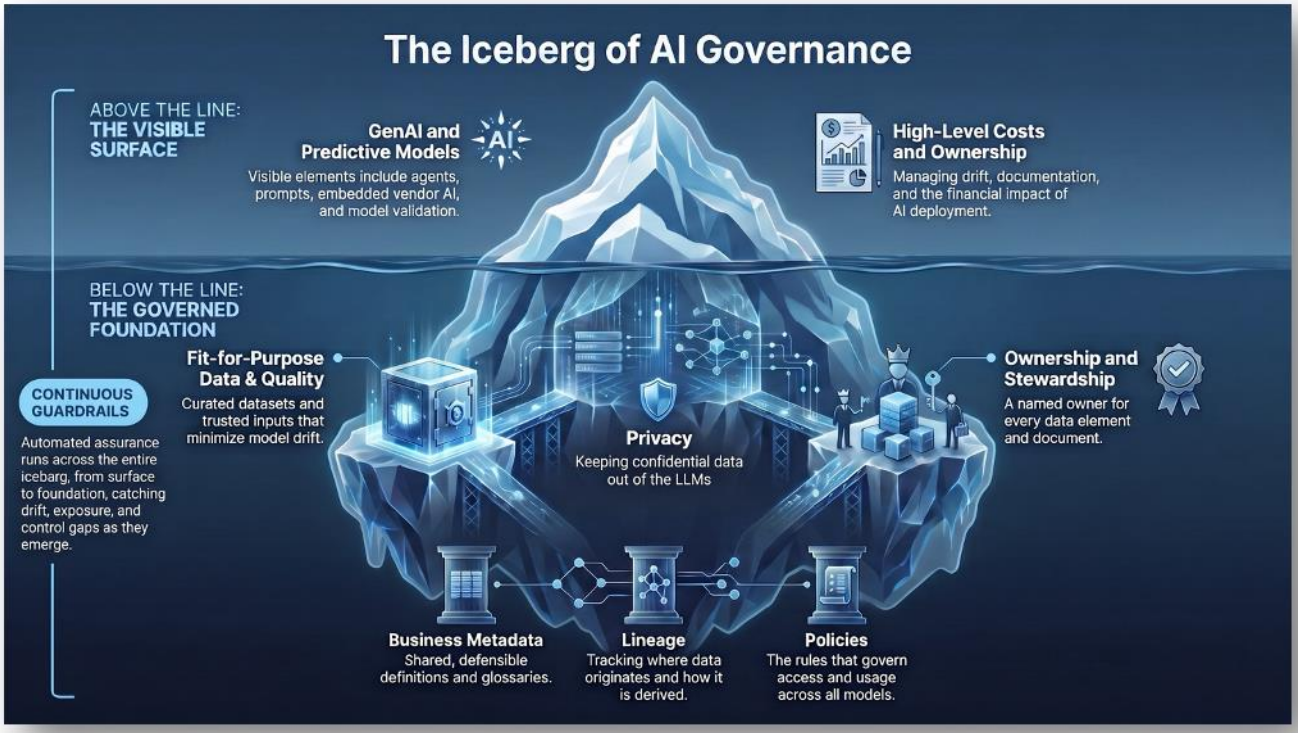
At 200 agents: the appearance of governance

Control plane rule

- Applies to every agent
- Coverage is independent of volume
- Evidence produced as a by-product

At 200 agents: a governance function

The economics make this decisive rather than preferable. A review board costs more as volume rises. A platform control costs once and covers everything after. At ten use cases that is an inconvenience. At two hundred agents it is the difference between a governance function and the appearance of one.



DO NOT GOVERN AGENTS AND MODELS WITHOUT GOVERNING THE DATA AND POLICIES UNDERNEATH THEM.

Fit-for-purpose data and quality Curated datasets and trusted inputs that minimize drift	**Privacy** Confidential data kept out of the LLMs
Business metadata Shared, defensible definitions and glossaries	**Lineage** Where data originates and how it is derived
Ownership and stewardship A named owner for every data element and document	**Policies** The rules that govern access and usage across all models

Most AI governance tools sit above the line, inventorying models and LLMs, tracking prompts and storing policies. That work is the tip. It cannot detect a confidential attribute reaching an LLM through an unclassified document, or a dataset that was never fit for the decision the model now makes. Auditrol's connected governance maps every model and agent to its identity and its access to specific data attributes and documents, so continuous guardrails run across the whole iceberg and catch drift, exposure and control gaps as they emerge.

REFERENCE MODEL

A WORKING REFERNCE MODEL

BE THE RAIL, NOT THE CASUALTY

> *The design rests on one discipline: reasoning is separated from execution. Agents interpret intent and plan; deterministic workflows carry out approved actions. Between them sits a control plane that enforces policy, logs every step and requires approval for anything material.*[8]

GOVERNED GENAI REFERENCE ARCHITECTURE

1. Presentation	Where the request is made: copilot, code assist, agent console.
2. Control plane	Policy enforcement, scanning, redaction, entitlement inheritance, human approval gates. Every request passes here.
3. Reasoning	Agents interpret intent and produce a plan.
4. Context	Approved catalog and document index, filtered by the requesting user's entitlements.
5. Execution	Deterministic workflows carry out the approved plan.
6. Governance	Immutable logs, metering, quality signals, dashboards and governance packs.

The choice that does the most work is narrow: agents hold no database credentials and no network path to data systems, so a compromise or hallucination cannot read or alter data. The assurance layer turns operation into evidence: immutable logs, cost metering by agent and use case, and quality signals for failures, refusals and overrides.

EXECUTIVE TAKEAWAY

Judge any reference architecture. Can a non-engineer reconstruct what a given agent did last quarter, and can you show the control that stopped it doing more? If the answer needs an engineering ticket, the architecture is not yet producing assurance.

ACTION

WHAT TO DO NOW

None of this requires waiting for a prescriptive standard. It requires five decisions and a sequence.

1	**Write down the standard you intend to be held to.** Map it to NIST AI RMF and AI 600-1, approve it at the right level, and be ready to explain why it fits your risk profile.
2	**Build an AI register that is not your model inventory.** Every agent and AI-enabled workflow in production, development and pilot, with an owner, a permission scope, a blast radius, an autonomy tier and a tested stop.
3	**Govern the layer underneath before you expand the layer above.** Fit-for-purpose datasets, lineage, classification, named owners and the policies that govern access. No agent is safer than the data it reads.
4	**Move controls from documents into the control plane.** Encode what can be encoded and keep an explicit list of what could not be, with the reason. That list is your residual risk register.
5	**Match the control to the risk, not to the average.** Rules-based for narrow and static systems, outcome testing where the logic is opaque, propagation controls where a failure can travel.

PHASE #1	PHASE # 2	PHASE #3
Inventory	**Tier and triage**	**Close the evidence gap**
Find every AI system running, including consumer-tier tools. Assign an owner to each. Expect the count to exceed the estimate.	Classify by autonomy and breadth of impact. Put a tested stop on anything that can take a material action unaided.	For the top tier, establish logging a non-engineer could use to reconstruct a decision. Take the standard to the risk committee.

Every model and agent is only as sound as the data, documents and policies beneath it. Auditrol maps that layer and proves it holds.
Let's map yours: contact@auditrol.com.

Sources

1. SAS and Coleman Parkes Research, “Your Journey to a GenAI Future” (SAS Institute Inc., 2024), based on a global survey of 1,600 organizations across sectors; banking and insurance recorded the highest generative AI adoption. Figures cited here are for leaders in the highest-adopting sectors. sas.com.

2. Richard Fleming, Maria Teresa Tejada, Brendan O’Rourke, and David Allen, “Agentic AI Governance, Risk, and Controls for Business Leaders” (Bain & Company). bain.com.

3. National Institute of Standards and Technology, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, January 2023; and “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,” NIST AI 600-1, July 2024. nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.

4. Fintech Open Source Foundation (FINOS), AI Governance Framework: risk catalogue and validated control set. Twenty-three of the controls referenced here are FINOS-validated. air-governance-framework.finos.org.

5. Auditrol, Inc., “NIST AI RMF and AI 600-1 Control Overlap and Mitigation Mapping,” 2026. Unified catalogue of 40 distinct controls, of which 17 (42%) are common to AI governance and predictive model governance and 20 are AI-specific; 37 controls carry a named NIST subcategory reference (23 FINOS-validated, 14 Auditrol-recommended) against a register of 33 risks; requirement coverage six covered, six bridged, no open gaps.

6. Gianvito Lanzolla, Margherita Pagani, and Christopher L. Tucci, “Scaling AI With Adaptive Governance,” MIT Sloan Management Review, Summer 2026; based on 2022 to 2025 interviews at Microsoft, Barclays, Kyriba, Nasdaq, Lloyds Bank, Danske Bank, and the Abu Dhabi Department of Finance. https://doi.org/10.63383/lVMh6418.

7. ISO/IEC 42001:2023, “Information technology. Artificial intelligence. Management system,” referenced for third-party lifecycle responsibilities and data classification expectations. iso.org.

8. Auditrol, Inc., “Governed GenAI Reference Architecture,” 2026.