THE READINESS CHASM:
TRANSLATING AI GOVERNANCE STRATEGY INTO PRODUCTION-READY INFRASTRUCTURE

A Policy Analysis

Joshua Reh | Cijara Group LLC | July 2026

THE GOVERNANCE GAP THAT THE NUMBERS REVEAL

Enterprise AI governance has a measurement problem, and the numbers make it visible. Seventy percent of organizations have established cross-functional AI oversight committees. Sixty-seven percent are actively upgrading enterprise systems. Forty-eight percent have operational AI governance guardrails in progress. Forty-one percent have established dedicated AI governance teams. Yet only 14 percent of organizations have achieved full preparedness to deploy AI capabilities at scale.

This is not a resource problem or an intent problem. It is a translation problem. Governance is advancing on paper. Production-ready infrastructure remains critically underdeveloped. Committees exist; controls do not. Frameworks are adopted; implementations are not. The gap between the 70 percent who have established oversight structures and the 14 percent who can actually deploy with confidence is what this paper calls the readiness chasm, and closing it requires a different kind of analysis than the field has typically produced.

Most governance literature addresses the policy layer: which frameworks apply, which regulations matter, which principles should guide AI deployment. That work is necessary. It is not sufficient. What organizations need -- and what governance research has underproduced -- is the operational translation layer: the specific architectural decisions, control mechanisms, and oversight structures that turn boardroom strategy into running infrastructure. This paper addresses that translation.

WHY AGENTIC AI REQUIRES A DIFFERENT GOVERNANCE ARCHITECTURE

The governance frameworks most organizations have built were designed for a different class of system. Traditional AI -- chatbots, recommendation engines, classification models -- is assistive. It generates outputs in response to prompts, and its primary risks are hallucination, bias, and misinformation. The primary controls are output filters and human proofreading. The system impact of a failure is low; it does not alter underlying databases or trigger downstream actions.

Agentic AI is categorically different. It is autonomous: it pursues objectives, chains decisions across multiple steps, calls external APIs, routes data between systems, and executes actions without per-step human instruction. Its primary risks are irreversible state changes, corrupted data, and unauthorized execution of consequential decisions. The system impact of a failure is severe; an unbounded agent can delete data, modify production databases, or violate regulatory requirements before any human observer detects what has happened.

The governance implication is direct: when AI moves from suggesting to doing, governance must shift from content filtering to action bounding. This is not an incremental adjustment to existing frameworks. It is a structural redesign of where controls sit, what they enforce, and who is accountable for them. Organizations that apply traditional AI governance architecture to agentic systems will find that their controls are positioned at the wrong layer of the stack to intercept the risks that actually matter.

THE SHADOW AGENT PROBLEM AND WHY OBSERVABILITY IS THE FOUNDATION

Before any technical safeguard architecture can function, an organization must know what it is governing. That condition is not currently met at most enterprises. Twenty-nine percent of employees are already using unsanctioned AI agents for work tasks. Thirty-five percent of organizations admit they could not shut down a rogue AI agent if one emerged. These are not edge cases. They are the baseline condition of the enterprise AI environment as agentic deployment accelerates.

The liability dimension is structural, not incidental. Agents inherit human permissions. An unsanctioned agent with broad tool-calling access is not simply a compliance failure; it is an active operational liability capable of executing the same adversarial failure modes as a sanctioned agent with insufficient boundary constraints -- but without any of the governance infrastructure that would detect or contain the damage.

The governance principle that follows is absolute: you cannot govern what you cannot observe. The first operational requirement of an agentic governance program is a centralized agent registry -- a single source of truth for all agents operating in the enterprise environment, distinguishing sanctioned agents from third-party integrations and flagging quarantined shadow agents. This registry is not a compliance artifact. It is the prerequisite for every other control in the architecture. Without it, risk-tiered autonomy limits, access controls, and observability telemetry all operate against an incomplete picture of the actual threat surface.

THE THREE-LAYER SANDBOX: PRACTICAL BOUNDARY ARCHITECTURE

With the observability foundation in place, the operational question becomes how to set boundaries that hold under adversarial conditions. The answer is a three-perimeter architecture, each layer addressing a different dimension of agentic risk.

The first perimeter is the model layer, addressing alignment. Prompt guardrails prevent policy violations and ensure data sanitization -- PII detection and masking -- before data leaves the enterprise environment. This is where content filtering belongs; it is the appropriate control for the output of the model, not the actions of the agent.

The second perimeter is the orchestration layer, addressing behavior. This is the layer most governance frameworks have not yet specified. Agentic systems can enter infinite loops in API calls, chain actions faster than human review can intercept, and pursue objectives through paths that no governance team anticipated at design time. The orchestration layer requires infinite loop detection, mandatory interruptability, hard-coded timeouts for looping actions, and a structural kill switch that can halt runaway processes without corrupting production data. Branch chain rollbacks -- the ability to attempt an action in a sandbox, fail safely, and roll back without contaminating the production environment -- are the technical implementation of the bounded cost principle: for autonomous automation to work at scale, the cost of failure must be strictly contained.

The third perimeter is the tool layer, enforcing least privilege. Role-Based Access Control restricts each agent to the specific APIs required for its defined function. An inventory management agent can read inventory; it cannot trigger purchase orders. An HR promotion agent can read the requesting employee's performance history; it cannot query the department's salary table. This maps directly to the SR 26-2 effective challenge principle: agents cannot independently execute consequential decisions. Autonomy must be designed, not assumed, and governance must be embedded into product architecture rather than bolted on post-launch.

RISK-TIERED AUTONOMY: THE PRACTICAL CALIBRATION FRAMEWORK

Not all agentic deployments carry the same consequence profile, and governance architecture that treats them identically will either over-restrict low-risk automation or under-protect high-risk decision-making. A risk-tiered autonomy framework calibrates controls to the actual stakes of the use case.

At the minimal risk tier -- internal productivity tools, spam filters, scheduling agents -- the agent drafts and executes low-stakes tasks, and the human applies the output. Full autonomy is appropriate here; the reversibility of any individual action is high and the consequence of a single failure is bounded.

At the limited risk tier -- customer service chatbots, research synthesis agents, document processing workflows -- the agent resolves standard cases independently but is hard-coded to require escalation for actions exceeding defined thresholds. A customer service agent that handles standard refund requests autonomously must be architecturally blocked from processing refunds above a defined dollar threshold without human authorization. Bounded autonomy is the appropriate posture.

At the high-risk tier -- credit decisions, HR actions, life-safety applications, and any use case falling under EU AI Act Annex III high-risk classifications -- explicit human approval is structurally required for every consequential action. Zero autonomy is not an operational constraint; it is a regulatory requirement that the architecture must enforce, not merely recommend. The system cannot rely on the agent's judgment about whether a situation requires human review. The threshold is pre-defined, hard-coded, and non-negotiable.

FROM HUMAN-IN-THE-LOOP TO HUMAN-IN-THE-LEAD

The traditional human-in-the-loop model positions the human at the end of the pipeline. The agent drafts an action, stops, and explicitly asks for approval before proceeding. This model is appropriate for high-risk, irreversible decisions -- it maps directly to the hard-stop zone of the risk-tiered framework above. But applied uniformly across all agentic workflows, it creates bottlenecks that defeat the operational value of the technology and produce human reviewers who are exhausted, rubber-stamping, and providing no meaningful

oversight.

The governance architecture that scales is Human-in-the-Lead. The human moves upstream: no longer checking every output, but instead defining strategic goals, designing the agent's operating environment, monitoring observability metrics, and managing overall system performance. The human's role shifts from output reviewer to system manager. This evolution has three phases. In Phase 1, humans are relieved of rote end-of-pipe output checking as low-risk automation matures, and the work moves upstream to designing compliance agents and specialized governance tooling. In Phase 2, as agents begin communicating with other agents in agent-to-agent (A2A) architectures, humans become system orchestrators rather than task operators. In Phase 3, oversight evolves to observing evaluations, monitoring the observability control plane, and attending to the underlying context of agent interactions rather than their individual outputs.

The policy implication is significant: workforce transformation for the agentic era requires defining exactly which decisions demand meaningful human oversight, consistency, and explainability, and building the organizational capability to exercise that oversight effectively at the system level rather than the transaction level.

TRANSLATING BOARDROOM STRATEGY INTO OPERATIONAL RISK MACHINERY

The readiness chasm described at the outset of this paper has a structural cause: boardroom AI governance strategy and operational risk infrastructure have been developed in parallel rather than as an integrated system. The synthesis that closes the gap maps KPMG's five governance pillars directly onto the NIST AI RMF's functional categories.

Strategic oversight and board work -- KPMG's P1 and P5 -- dictates the NIST Govern function, establishing organizational culture, baseline rules, and regulatory compliance posture. This is where the AI governance charter, the cross-functional council, and the board's AI oversight mandate live. Security -- KPMG's P2 -- operationalizes into NIST's Respond and Manage functions, where active threat detection, vendor dependency management, and the structural controls against data poisoning, agentic misalignment, and shadow AI operate. Workforce accountability -- KPMG's P3 -- also feeds the Manage function, enforcing human-in-the-loop controls and defining the non-negotiable judgment threshold below which AI output cannot substitute for human decision-making. Trustworthy AI -- KPMG's P4 -- drives NIST's Map and Measure functions: quantitative evaluation of fairness against protected classes, privacy controls across the data supply chain, and the continuous monitoring infrastructure that makes performance degradation visible before it becomes a regulatory event.

Risk cannot be measured until the precise context of the system is mapped. Visibility across the AI pipeline is highly fragmented; builders, operators, and end-users rarely see the complete end-to-end system. The tolerance for risk is highly variable: the acceptable risk for an internal HR tool is categorically different from an autonomous trading algorithm. The mapping exercise that connects boardroom strategy to operational controls must be done at the use-case level, not the enterprise level. A governance framework that applies the same controls to every AI deployment regardless of its consequence profile is neither proportionate nor defensible under existing regulatory standards.

NAVIGATING THE FRAGMENTED GLOBAL REGULATORY LANDSCAPE

Organizations operating across jurisdictions face a regulatory environment that is fragmented by design philosophy rather than merely by implementation timing. The EU's approach is risk-based and rights-centered, with a centralized, binding Act that categorizes societal risk across all sectors and imposes strict compliance obligations including outright bans on certain high-risk use cases. The compliance burden is real; the legal certainty is also real.

The United States approach is innovation-first, relying on voluntary compliance and agency directives rather than binding statutory requirements. The flexibility is genuine; so is the risk of losing global regulatory influence to frameworks being written elsewhere. The executive order landscape evolves faster than enterprise governance programs can track, creating operational uncertainty that conservative compliance teams interpret as a reason to wait rather than a reason to build durable internal standards.

For practitioners and policymakers alike, the strategic response to regulatory fragmentation is not to wait for harmonization that may not arrive on any useful timeline. It is to build governance infrastructure anchored to the most demanding applicable standard -- currently the EU AI Act for high-risk classifications -- while designing for the adaptability that less prescriptive regimes require. An organization that can demonstrate compliance with the EU AI Act's transparency, bias testing, and human oversight requirements for high-risk AI can

credibly demonstrate compliance in any jurisdiction. The reverse is not true.

THE 90-DAY PATH FROM STRATEGY TO OPERATIONAL READINESS

The readiness chasm does not close through additional strategy documents. It closes through sequenced operational investments. A 90-day implementation arc provides a practical roadmap.

The first 30 days are dedicated to inventory and evaluation. This means deploying an evaluation harness -- tools such as the UK AI Security Institute's inspect_ai provide policy-grade agent observability and bias testing -- and establishing the centralized registry that baselines active and shadow agents across the enterprise. Without this inventory, every subsequent investment operates against unknown scope.

Days 31 through 60 focus on sandboxing and workflow governance. A reference agent is installed in a sandboxed environment with full branch chain rollback capability, and a cross-functional AI Governance Council is established with the mandate to define risk tiers, autonomy limits, and approval workflows for all production deployments. This council is not an advisory body. It is the decision-making authority for the risk-tiered framework, and its decisions must be implemented architecturally, not honored voluntarily.

Days 61 through 90 operationalize continuous observability. Vulnerability scanning is integrated into the production monitoring stack, real-time security telemetry is tied to business KPIs rather than isolated IT metrics, and structural kill switch capabilities are enforced for all production agents. At the end of 90 days, the organization has not eliminated AI risk. It has built the infrastructure that makes AI risk visible, measurable, and manageable. That is the condition that separates the 14 percent from the 86 percent.

IMPLICATIONS FOR GOVERNANCE RESEARCH

The analysis above identifies several areas where the existing governance research agenda has not yet produced the operational specificity that practitioners require.

First, the risk-tiered autonomy framework described here is practitioner-developed rather than research-validated. There is no published empirical work establishing where the calibrated boundaries between full autonomy, bounded autonomy, and zero autonomy should sit for specific use case categories. The thresholds used in practice are judgment calls made by individual governance teams without a shared evidentiary base.

Second, the Human-in-the-Lead evolution described in Phase 2 and Phase 3 -- where agents communicate with other agents and humans become system orchestrators -- has no established governance standard. Agent-to-agent architectures introduce accountability gaps that no current framework addresses: when an orchestrating agent delegates a task to a subordinate agent that causes harm, existing liability frameworks do not clearly assign responsibility. This is an open problem with significant near-term practical consequence.

Third, the Machine Learning Bill of Materials -- the requirement to log the exact lineage, training data, and vendor dependencies of every deployed model -- is moving from voluntary standard toward 1LOD and 2LOD control requirement. The research infrastructure for evaluating ML-BoM compliance at enterprise scale does not yet exist. Practitioners are building audit artifacts without agreed standards for what those artifacts must contain or how they will be evaluated.

Governance is not a brake on AI deployment. It is the structural integrity that allows organizations to move toward the frontier of AI capability with ambition rather than fear. The work of closing the readiness chasm -- connecting boardroom strategy through operational architecture to running production controls -- is where practitioners with institutional risk management experience and hands-on deployment exposure have the most direct contribution to make to the research agenda. That translation has barely begun.

Joshua Reh is Principal and co-founder of Cijara Group LLC, a federal subcontracting firm specializing in AI governance, program management, and regulatory compliance advisory. He holds 15+ years of enterprise risk, model risk, and operational risk experience across Goldman Sachs, Bank of America, and Wells Fargo. Contact: Joshua@CijaraGroup.com | cijaragroup.com