SECURING THE AUTONOMOUS VAULT:
PII GOVERNANCE IN THE ERA OF AGENTIC AI

A Policy Analysis

Joshua Reh | Cijara Group LLC | July 2026

THE GOVERNANCE GAP ENTERPRISES CANNOT AFFORD TO IGNORE

Enterprise AI adoption has reached an inflection point. Seventy-four percent of enterprises plan to deploy agentic AI systems within two years to autonomously execute workflows. Yet only 21 percent of those organizations possess a governance model mature enough to manage the risks those systems create. The consequence is predictable: 80 percent of enterprise AI projects remain trapped in pilot phases, not because the technology fails, but because the governance infrastructure needed to deploy it responsibly does not yet exist.

This is not a technology problem. It is a governance architecture problem. And the cost of solving it incorrectly -- or not solving it at all -- is not merely regulatory. It is institutional.

Agentic AI systems are categorically different from the generative AI tools that preceded them. A generative system is passive; it responds to prompts and produces outputs. An agentic system is active; it pursues objectives, plans multi-step workflows, calls external APIs, routes data between systems, and triggers consequential events without per-action human instruction. The core security risks shift accordingly. Where generative AI exposes organizations to hallucination and prompt-based data leakage, agentic AI introduces autonomous database modification, unauthorized API execution, and what this paper terms the "Double Agent" vulnerability -- prompt injection attacks that redirect a trusted agent's behavior from the inside.

When those agents operate inside enterprise HR platforms like Workday, where they interact with Social Security numbers, salary histories, student performance evaluations, and medical leave records, the stakes are not abstract. They are immediate, material, and regulatory.

THE GROUNDED SCENARIO: THE UNIVERSITY HR VAULT

To make the governance challenge concrete, consider a representative deployment scenario. A university deploys an AI agent integrated with Workday to automate HR workflows, including processing a faculty member's promotion and salary adjustment. The agent must access enough personally identifiable information (PII) to evaluate promotion criteria -- performance history, compensation tier, departmental benchmarks -- while being systematically prevented from exposing peer salary data, student records, or executing the final compensation change without explicit human authorization.

This scenario is not hypothetical. It represents a class of deployment that dozens of higher education institutions and enterprise HR functions are actively building or evaluating. It is also a scenario where the gap between what the agent is capable of doing and what it is permitted to do is the entire governance problem.

The vulnerability that most organizations underestimate is not the agent behaving badly in obvious ways. It is the "Double Agent" failure mode: a malicious actor submits a heavily obfuscated adversarial prompt disguised as a routine HR inquiry. The agent, lacking strict boundary constraints, accepts this input as a priority objective and overrides its original safety instructions. The over-permissioned agent then leverages its Workday API access to extract restricted student PII and faculty performance reviews and returns that data to the unauthorized requester -- bypassing standard perimeter defenses because the agent itself is a trusted entity in the system.

This failure mode does not require a sophisticated attacker. It requires only an agent with excessive permissions and insufficient boundary enforcement. Both conditions are common in early-stage enterprise deployments.

WHY DATA GOVERNANCE ALONE IS INSUFFICIENT

A common organizational response to this problem is to treat it as a data governance issue. That response is partially correct and dangerously incomplete.

Data governance addresses inputs: how data is managed, protected, and made reliable before it reaches a model. It asks whether the dataset is accurate, who owns it, and whether the organization is compliant with GDPR and HIPAA retention requirements. These are necessary questions. They are not sufficient ones.

AI governance addresses outcomes: how AI systems are managed, kept safe, and constrained during operation. It asks whether the system's outputs are fair and robust, how the organization handles model drift or degraded performance, and who holds accountability for human oversight when the system acts autonomously. The target risks are different -- biased outputs, autonomous misuse, and opaque decisions with consequences for real people -- and they require different controls.

Organizations that treat AI governance as an extension of data governance will systematically underinvest in the outcome-side controls that agentic systems require. The result is a governance model that protects the vault's contents but leaves the vault's door open.

TRANSLATING REGULATORY FRAMEWORKS INTO OPERATIONAL CONTROLS

Three regulatory frameworks currently define the outer boundary of what responsible agentic AI deployment requires. Each maps to specific operational obligations that organizations can and must implement.

The EU AI Act establishes risk-tier classification and Fundamental Rights Impact Assessments for high-risk AI use cases. In the university HR scenario, the Workday Salary Agent is classifiable as high-risk under the employment decision category, legally mandating strict transparency, bias testing, and human oversight before deployment. This is not a future obligation. For organizations operating in or selling to EU markets, it is a present one, with high-risk compliance obligations deferred to December 2027 under the Digital Omnibus amendments.

The NIST AI Risk Management Framework requires organizations to Govern, Map, Measure, and Manage AI risk across the full system lifecycle. Applied to the HR agent, this means mapping the agent's data provenance, measuring its performance for demographic bias against protected classes, and managing its API tool-calling permissions as a first-order governance artifact rather than a technical configuration detail. The NIST AI RMF is the most operationally tractable of the three frameworks for organizations with existing enterprise risk infrastructure; its functional categories map directly onto second-line-of-defense risk management processes.

ISO/IEC 42001 establishes the requirements for certifiable AI management systems, including continuous audit trail obligations. In practice, this means that every time the HR agent accesses Workday PII, that access must be logged, the log must be immutable, and the record must be available for regulatory review on demand. This is not meaningfully different from what financial institutions have been required to maintain for model risk under SR 26-2; the underlying governance logic is the same, applied to a newer class of system.

THE TECHNICAL SAFEGUARD ARCHITECTURE

Regulatory compliance defines the boundary. Technical architecture defines whether the boundary holds under adversarial conditions. Three safeguards are necessary and, taken together, sufficient to secure agentic PII workflows against the failure modes described above.

Zero Trust Identity and Access. The agent must not be granted broad system permissions at deployment and trusted thereafter. Every access request must be verified against three conditions: cryptographic identity confirmation, current risk level, and behavioral baseline. In the university scenario, this means the agent is permitted read-only access strictly to the requesting professor's historical performance data required for the promotion summary. Any attempt to read the entire department's salary table is denied, logged as a violation, and triggers temporary quarantine of the agent's privileges. Trust is not a deployment-time decision. It is a continuous verification process.

Data Loss Prevention at the Output Layer. The agent's internal access permissions address what data the system can retrieve. DLP controls address what data can leave the secure environment in the agent's outputs. All university documents and Workday records must be pre-tagged by classification tier (Secret, Confidential, Internal, External), with the agent hard-coded to ignore Secret-tagged records such as medical leave files. As the agent formulates its promotion summary, output passes through a real-time DLP engine that detects and masks raw PII -- converting a Social Security number to XXX-XX-1234 -- before the output reaches the user interface. The sensitive data never leaves the secure environment regardless of what the agent retrieves.

Observability and the Control Plane. Neither access controls nor DLP are sufficient without an observability layer that provides continuous visibility into what agents are actually doing. This requires three components operating together: a central agent registry that distinguishes sanctioned agents from third-party integrations and flags quarantined shadow agents; a data lineage tracker that records

every read and write operation with the agent identity, timestamp, and data classification; and security telemetry that generates real-time alerts for anomalous behavior, including infinite loops in API calls and unauthorized write-access attempts to protected databases. Shadow AI -- unsanctioned agents deployed by employees outside IT oversight -- is not primarily a technical problem. It is an observability problem. You cannot govern what you cannot see.

THE HUMAN-IN-THE-LOOP FRAMEWORK

Technical safeguards address what the system can do. The human-in-the-loop framework addresses what requires human judgment before the system acts. These are not the same boundary, and conflating them produces governance models that are either too restrictive to be useful or too permissive to be safe.

A workable HITL framework distinguishes three operational zones based on consequence severity. In Zone 1, the agent operates with full autonomy for low-consequence, reversible actions -- drafting a summary of public promotion guidelines, for example -- and proceeds without interruption. In Zone 2, the agent executes higher-consequence tasks -- querying Workday, reading performance records, drafting a promotion justification memo -- but the output requires human review and verification before any further action. This zone addresses hallucination risk: the memo may be structurally correct but factually wrong about specific data points, and a human manager must verify accuracy before the memo is transmitted. In Zone 3, the agent encounters an irreversible action -- executing a salary tier change directly in the Workday database -- and is frozen. A mandatory, cryptographically-signed approval request is sent to the HR Director, and the database write cannot execute without explicit human authorization. The system architecture enforces this; it is not a policy aspiration that depends on the agent's judgment.

The HITL framework is the governance mechanism that most directly addresses the concern that the loss of human oversight as AI systems grow more capable is the central risk that outpaces any technical safeguard. The HITL framework does not slow agentic AI. It creates the conditions under which agentic AI can be deployed at scale with defensible oversight.

THE LAYERED DEFENSE MODEL AND ITS IMPLICATIONS FOR POLICY

The three safeguard layers described above do not operate independently. They form a single governance architecture: a policy layer that defines the agent's approved scope and legal boundary; a technical layer that cryptographically enforces those boundaries at the access, output, and identity levels; and a human layer that provides the final failsafe for consequential, irreversible actions.

What this architecture makes visible is that enterprise AI governance is not a compliance exercise. It is the infrastructure that allows organizations to deploy autonomous intelligence at scale without scaling the risk proportionally. Organizations that invest in this architecture accelerate deployment by eliminating the case-by-case negotiation that currently traps 80 percent of AI projects in pilot. They build scalable trust with regulators and internal stakeholders through observable, auditable systems. And they reduce liability by intercepting the failure modes -- unauthorized data access, prompt injection, shadow AI accumulation -- that produce costly regulatory actions and reputational damage.

For governance researchers and policymakers, the key implication is structural: the governance frameworks that currently exist -- NIST AI RMF, EU AI Act, ISO/IEC 42001 -- provide the right conceptual architecture but insufficient operational specificity for agentic deployments. The HITL zone framework described above is one example of an operational construct that existing guidance does not yet specify. The agent registry and shadow AI observability requirements are another. The field needs practitioner-developed operational standards that translate framework obligations into implementable controls, tested against real deployment scenarios rather than hypothetical ones.

That translation work -- from regulatory intent to enforceable operational control -- is where practitioners with institutional risk experience and hands-on deployment experience can contribute most directly to the governance research agenda. It is also where the gap between existing guidance and enterprise readiness is largest.

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