AI GOVERNANCE SERIES | SECURITY PERSPECTIVE

AI Governance Is Becoming an AI Security Problem

Most Organizations Are Defending AI Systems with Tools Designed for a Different Threat Model.

By Greg Aldrich | Global CIO & Strategic Advisor | May 2026

In recent months, I have discussed AI governance as an operating model issue: the gap between governance intent and execution, and between policy documents and enforceable controls. This perspective remains accurate.

However, I have previously underemphasized the security aspect of governance, which recent events have made critical to address.

AI governance without AI-native security is not true governance; it is simply aspirational.

The evidence is accumulating quickly. HiddenLayer's 2026 AI Threat Landscape Report, published in March, found that autonomous agents now account for more than one in eight reported AI breaches as enterprises move from experimentation to production. Prompt injection attacks caused an estimated $2.3 billion in losses globally in 2025. OWASP ranked prompt injection as the single highest-severity vulnerability category for deployed language models — above data poisoning, above model theft, and above insecure output handling. And current detection tools catch only 23% of sophisticated injection attempts.

Most organizations developing AI administrative frameworks are not dealing with these security challenges. They create policies and form committees yet deploy AI systems versus threats their current security tools cannot manage.

> ***"Governance without observability is policy theater. And most organizations are still running the theater."***

Traditional Security Was Not Designed for AI Systems

Current enterprise security models were designed for deterministic software: predictable applications, validated inputs, and enforceable boundaries at the network and identity layers. While not incorrect, this approach is insufficient for AI.

AI systems vary greatly in ways that impact security. They interpret natural language instead of executing code, cannot consistently distinguish operator instructions from processed content, and make probabilistic decisions influenced by unforeseen inputs. As these systems transition from answering questions to taking actions such as executing code, sending emails, or modifying databases, the operational impact of a compromised AI agent becomes far greater than that of a compromised chatbot.

The vulnerability surfaces that AI introduces are genuinely new:

Model inference attacks — probing a model's behavior to extract information about its training data or architecture without direct access

Prompt injection — embedding malicious instructions in content the AI will process, causing it to override its system instructions and execute attacker-directed actions

Model supply chain compromises including poisoning model artifacts, embeddings, or weights before deployment

Training data contamination — manipulating data used to fine-tune models to embed backdoors or biases

RAG pipeline poisoning — corrupting the vector databases and document stores that retrieval-augmented systems depend on

Agent tool poisoning — manipulating the tool definitions that govern what actions an AI agent can take and when

Traditional web application firewalls, endpoint detection tools, and identity and access management platforms do not deal with these vulnerability points. Effective protection requires security controls that understand AI systems, not just networks and applications.

> ***"Prompt injection is not a chatbot trick. It is the new phishing, and unlike phishing, it requires no human to click anything."***

The Threat Is Not Theoretical

I want to be precise here, because AI security can easily tip into either hype or dismissiveness. The threat is real, documented, and accelerating.

In 2026, a zero-click attack against a major enterprise AI assistant worked as follows: an attacker sent a crafted email to anyone in the target organization. When any user later asked the AI assistant

a question, it retrieved the poisoned email, executed the embedded instructions, and exfiltrated sensitive data via an image URL, all without a single click from the victim. No malicious link. No anomalous login. No binary to detonate. The AI did the attacker's work, autonomously, as part of its normal operation.

Earlier this year, researchers documented a vulnerability in GitHub Copilot, CVE-2025-53773, where hidden prompt injection in pull request descriptions enabled remote code execution, with a CVSS score of 9.6. The EchoLeak vulnerability found in Microsoft 365 Copilot demonstrated that a zero-click prompt injection could silently access and exfiltrate enterprise data.

HiddenLayer's research team has documented similar patterns specifically in agentic systems: MCP tool poisoning, where faint variations in tool descriptions cause agents to select tools that expose sensitive data or perform unauthorized operations. The agent cannot verify tool intent independently. It trusts the schema as authoritative. An attacker who understands this can manipulate what the agent does without ever touching the model.

These incidents are not anomalies. They are the predictable result of deploying systems that analyze and act on natural language within environments in which adversaries comprehend those processes.

HiddenLayer as a Market Signal

To clarify, I am not reviewing HiddenLayer as a product. Instead, I reference its existence and the category it represents as signs of the direction of enterprise AI governance.

HiddenLayer was founded in 2022 by a team with deep roots in security and machine learning, motivated by a real-world adversarial ML attack experienced at Cylance. It is Gartner-recognized, backed by Microsoft's M12 venture fund, IBM Ventures, Booz Allen Ventures, and Capital One Ventures, and has raised $56 million. It has disclosed 48 CVEs in ML frameworks including PyTorch and TensorFlow and holds 25 granted patents regarding adversarial detection and model protection.

Its platform covers four areas that traditional security tools do not: AI asset discovery throughout environments; AI supply chain security, evaluating models and artifacts before deployment; AI attack simulation, continuously testing systems for vulnerabilities using attack techniques; and AI runtime security, monitoring models in production to detect and stop attacks in real time.

The strategic importance lies not in individual features, but in the recognition that AI systems require security controls throughout their entire lifecycle, from discovery to supply chain to runtime. These controls must be AI-native, rather than adapted from existing infrastructure security models.

The emergence of this category signals a shift in the market. When a well-funded, Gartner-recognized company with strong technical expertise develops a dedicated AI security platform, it reflects real enterprise demand driven by genuine threats that existing tools do not address.

> ***"The existence of AI-native security vendors is evidence that enterprises are recognizing traditional controls are insufficient. The market is telling you something. It is worth listening."***

The Governance Questions That Actually Matter

For CIOs and enterprise leaders, security discussions should focus on governance questions rather than threat classifications. The following are key questions for organizations that believe their current security posture addresses AI:

- What AI assets are present in your environment, and can your current security tools discover all of them? Most organizations cannot answer this confidently. Shadow AI—tools and agents deployed without IT oversight—is often the root cause of AI security failures.
- Do your security controls extend to the model layer, or only to the application and network layer? Traditional tools see inputs and outputs at the application boundary. They do not see what is happening inside model inference, inside embedding retrieval, or inside agent tool selection.
- Who is operationally responsible for AI runtime security? Most organizations have not defined this. In many enterprises, responsibility falls between teams, resulting in a lack of clear ownership.
- Are your AI systems' supply chains assessed before deployment? Models downloaded from public repositories, fine-tuned on third-party datasets, or integrated through external APIs carry supply chain risk that traditional software composition analysis tools are not designed to evaluate.
- What is your incident response model for AI security events? A prompt injection attack that leads to data exfiltration requires a response plan distinct from traditional ransomware responses. Most organizations have not developed such a plan.
- Does your governance framework distinguish between AI that answers and AI that acts? Generative AI that produces output for human review carries a different risk profile than agentic AI that executes actions autonomously. Most governance policies do not make this distinction explicit.

These questions do not require advanced tools, but rather governance clarity, which most AI governance frameworks lack.

The Organizational Reality

The greatest challenge in AI security is not technical, but in assigning clear accountability.

Most enterprises are organized around technology boundaries that predate AI: a security team that owns network and endpoint defense, a data team that owns data governance, an AI or innovation team that owns model deployment. None of these teams fully owns AI security, and each has reasonable grounds to argue it falls primarily in someone else's domain.

This creates a structural governance gap. Organizations are not neglecting AI security out of indifference, but because ownership models, responsibility frameworks, and necessary funding have not been explicitly established.

As discussed in Part 3 of this series, governance that exists only in the gaps between functions is ineffective. It results in risk accumulation despite good intentions.

The immediate need is not a new security platform, but a clear decision regarding ownership:

- Who has accountability for AI asset discovery and inventory?

- Who reviews models and AI artifacts before deployment?
- Who monitors AI system behavior in production?
- Who owns the AI incident response playbook?
- Who approves new AI tools and agent configurations?

Till these questions have clear, accountable answers, security tools address problems the organization has not formally committed to managing.

> ***"AI security is becoming a board-level governance issue, not because the board needs to understand adversarial ML, but because they need to understand that their AI investments carry risks their current security posture was not designed to manage."***

Where This Fits in the Governance Architecture

For readers of this series, it is important to clarify where AI security fits within the governance architecture I have outlined.

In the five-layer model I published recently, the assurance and evidence fabric runs cross-cutting across all layers. AI security, the capability to discover, monitor, test, and protect AI systems from AI-native threats, belongs primarily in that fabric and in the AI-native control layer. It is the enforcement infrastructure that makes the other governance layers defensible when something goes wrong.

TowerIQ provides portfolio visibility: what AI exists, what it costs, and what risk it creates. That visibility depends on accurately discovering AI assets, which requires AI-native discovery capabilities, not just CASB plus shadow IT detection. HiddenLayer's AI Discovery module addresses exactly that gap.

OutcomeOps provides organizational intelligence enforcement at the engineering layer: ensuring AI-generated code and decisions comply with corporate standards and constraints. That enforcement is most credible when the AI systems doing the generating are themselves secured against the prompt injection and supply chain attacks that might compromise their outputs.

The AI security layer does not replace these components; it enables their trustworthiness.

What Good Looks Like

Organizations leading in this area share common practices. They do not demand a comprehensive AI security platform initially, but they all begin with a governance decision.

- They inventory all AI assets, including shadow AI, and assign ownership for each. They track what is running, who approved it, and what data it accesses.
- Their security review process includes AI-specific checks such as model provenance, training data lineage, inference behavior amid adversarial inputs, and supply chain dependencies.

- They treat prompt injection as a primary threat in AI system design, not as an afterthought. This requires implementing architectural controls such as separating system instructions from user inputs, limiting agent tool permissions, and sandboxing agent actions, rather in place of relying solely on detection.
- They have established a distinct incident response playbook for AI security events, separate from those for application and infrastructure incidents.
- They assign operational ownership of AI runtime monitoring to a specific function, with clear accountability and dedicated budget.

Although these measures do not address every AI security risk, they distinguish organizations that proactively govern AI from those that rely on existing security teams to react when issues arise.

Final Thought

Throughout this series, I have argued that governance must shift from policy documents to operating models, from committee oversight to embedded controls, and from reactive review to proactive specification. The AI security dimension is the most urgent aspect of this principle.

The threat is present now. Autonomous agents account for over one in eight AI breaches. Prompt injection is the leading vulnerability in production AI deployments, and detection tools identify fewer than one in four sophisticated attacks. Most organizations are responding with governance frameworks suited to a previous era of technology.

HiddenLayer is notable not solely as a product, but as evidence that the market recognizes a problem most governing frameworks have yet to address. Organizations that succeed with AI in the coming years will not only deploy it effectively, but also govern it comprehensively, including the security dimension that many still overlook.

> ***"AI governance without AI-native security is aspiration, not architecture. The organizations that recognize that distinction now will be significantly better positioned than those that discover it after an incident."***

AI Governance Series: *Part 1: The AI Governance Tightrope | Part 2: The New Control Layer | Part 3: The Operating Model for AI Governance | Three Frameworks | Target Architecture | OutcomeOps: The Organizational Intelligence Layer*

About the Author

Greg Aldrich is a Global CIO and Strategic Advisor with 30+ years of experience helping boards, executive teams, and C-suite stakeholders navigate IT strategy, AI governance, and digital transformation. He serves as Senior Strategy Advisor at Blue Tree Technology Group and Senior Transition Architect at SDS Consulting, and currently serves as CIO at Andrew Wommack Ministries & Charis Bible College, where he chairs both the AI Committee and IT Steering Committee. He has advised organizations across financial services, gaming, healthcare, higher education, logistics, and nonprofit sectors.

Connect: linkedin.com/in/galdrich