August 2026

ICIT

When Decisions Have No Assurance, AI Risks Everything.

Governance in the Age of AI

Author: Malcom Harkins
www.icitech.org

Table of Contents

I. Executive Summary	3
II. The Bedrock of Governance Hasn't Changed, But the Stakes Have	4
III. AI Governance: The Missing Link Across Disciplines	4-5
IV. From Decision Rights to Decision Assurance	5-6
V. The Risk and Control Continuum: From First Line to Last	7-8
VI. Risk, Decisions, and Accountability	9
VII. The Rise of Agentic AI: A Governance Inflection Point	9-10
VIII. Designing Governance for an AI-Driven Enterprise	10
IX. The Future of Governance: Precision and Confidence	11
X. Final Analysis: AI Governance Is a Competitive Advantage	11
XI. Bibliography	13

About ICIT

The Institute for Critical Infrastructure Technology (ICIT) is a nonprofit, nonpartisan, 501(c)3 think tank with the mission of modernizing, securing, and making resilient critical infrastructure that provides for people's foundational needs. ICIT takes no institutional positions on policy matters. Rather than advocate, ICIT is dedicated to being a resource for the organizations and communities that share our mission. The views and opinions expressed in this essay are solely those of the author(s) and do not necessarily reflect the official policy or position of ICIT. Any assumptions made within the analysis are not reflective of the position of any entity other than the author(s). To learn more, please visit www.icitech.org

The Institute for Critical Infrastructure Technology

I. Executive summary

This paper is a companion to When Trust Has No Security, AI Risks Everything, which argues that AI cannot be trustworthy unless it is operationally secure, and asks whether an AI system can withstand attack once it is live. This paper asks the other half of the question: whether an organization can prove that the decisions its AI systems make, influence, or execute are sound, accountable, and controlled. The two questions are inseparable, but each paper is written to stand on its own.

Decision-making has always been the core function of governance. What has changed is who, or what, is now making the decisions. Artificial intelligence, and particularly agentic AI, has moved from supporting human judgment to making, influencing, and executing decisions across the enterprise, often at machine speed and machine scale. Traditional governance is built around decision rights: who decides, who approves, how escalation works. That work still has to happen, and it now covers less ground than it is used to. In an AI-driven enterprise, governance has to extend into decision assurance, meaning the organization can demonstrate with evidence that every decision, human or machine, falls inside defined risk tolerances, passes through effective controls, and produces the outcomes it was meant to produce.

This paper covers:

- Why AI collapses the traditional boundaries between corporate, IT, and cybersecurity governance
- The shift from decision rights to decision assurance, and how that shift tracks the "Well Built, Well Run, Trustworthy" maturity curve described in the companion paper's 9-Box of AI Trust & Security framework
- The risk and control continuum organizations need, from visibility and supply chain trust to validation and active defense
- Why agentic AI is a governance problem before it is a technology upgrade
- What governance design looks like in an AI-driven enterprise, and why decision assurance earns its keep commercially as well as with regulators

The central thesis is straightforward. Governance is still about decision-making, and it now has to account for machines that participate in, influence, or autonomously make those decisions. Governance sets the intent, and security proves the intent is being met. When decisions have no assurance, AI risks everything.

II. The Bedrock of Governance Hasn't Changed, But the Stakes Have

Decision-making is the component of governance that carries the most weight. Without clear decision rights, defined accountability, and a process that works, governance falls apart quickly. That has long been true in corporate governance and even more so in IT governance, where poorly coordinated decisions produce waste, higher costs, lost opportunity, and, as we have watched for decades now, rising cyber risk.
The principle holds; the environment it operates in has changed.

Artificial intelligence, and especially agentic AI, has changed how decisions get made, who or what makes them, how fast they execute, and how far their effects reach. Decisions are no longer confined to human actors working inside well-defined processes. They now sit inside systems, run automatically across workflows, and scale across whole organizations.

That leaves one hard requirement:
Governance is still about decision-making, but now it must account for machines that participate in, influence, or autonomously make those decisions.

III. AI Governance: The Missing Link Across Disciplines

Traditional governance has often been segmented:

- Corporate governance focuses on accountability, fiduciary responsibility, and shareholder outcomes
- IT governance aligns technology investments with business strategy
- Cybersecurity governance manages risk, resilience, and protection

AI fundamentally reshapes all three and inextricably commingles them. That entanglement creates a corporate governance exposure, and managing it takes a detailed understanding of AI risk.

This paper focuses on the links between them: the decision rights, accountability lines, and evidence requirements that let an organization govern AI responsibly. The companion paper referenced above takes up the inseparable question of whether the AI systems making those decisions can be trusted to behave as intended once they are live and under attack. Its argument is direct: AI cannot be trustworthy unless it is operationally secure. Unsecured trust isn't trust, it's negligence. Governance sets the intent, and operational security proves the intent is being met.

Without this integration, organizations face fragmentation:

- Business leaders make decisions without understanding technical risk
- IT deploys systems without full awareness of business impact
- Security teams mitigate threats without insight into decision logic

AI governance closes those gaps by keeping decision authority, accountability, and risk awareness aligned across the enterprise.

IV. From Decision Rights to Decision Assurance

Governance has traditionally defined:

- Who makes decisions
- Who can approve or veto them
- How escalation occurs

Those questions still need answers and answering them no longer finishes the job. Governance must move from defining decision rights to providing decision assurance.

Decision assurance means that an organization can:

- Demonstrate that decisions are made within defined risk tolerances
- Validate that controls governing those decisions are effective
- Provide evidence that outcomes align with intended objectives

Governance now must prove that decisions are correct, controlled, and trustworthy, on top of establishing who makes them.

Operational AI security follows the same maturity curve on the runtime side. The "9-Box of AI Trust & Security" framework describes it as a progression from Well Built to Well Run to Trustworthy: organizations cannot leap from building a model to trusting it. Decision assurance is the governance expression of that same discipline. Intent is not enough, and neither is building the thing correctly. Assurance comes from continuous, demonstrable proof that decisions, and the systems making or shaping them, behave as intended.

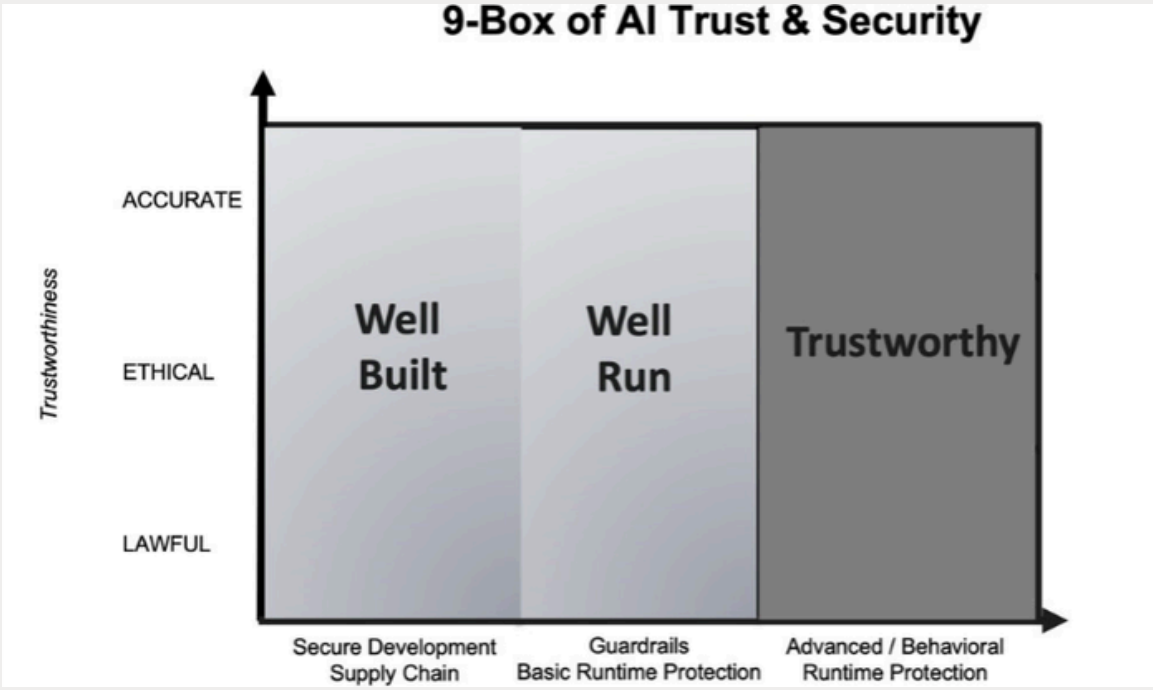

Figure 2. The 9-Box of AI Trust & Security: Well Built → Well Run → Trustworthy.

Organizations cannot leap directly from building a model to trusting it. Decision assurance is the governance-side expression of this same maturity progression. Source: Harkins, M., When Trust Has No Security, AI Risks Everything (HiddenLayer), adapted from Managing Risk and Information Security, 3rd Edition.

V. The Risk and Control Continuum: From First Line to Last

Effective governance depends on being able to demonstrate control across the whole lifecycle, from the first line of defense to the last. That includes:

1. Visibility: Knowing What Exists

You cannot govern what you cannot see. AI governance starts with a complete inventory of AI systems:

- Models (internal and third-party)
- Data sources
- Use cases and decision contexts
- Integrations into business processes

Without this visibility, governance becomes reactive and incomplete.

2. Supply Chain Trust: Understanding What You Depend On

AI systems are rarely built in isolation. They rely on:

- Pre-trained models
- External APIs
- Third-party data
- Embedded components

The result is a long AI supply chain. Governance must be able to evaluate:

- Model provenance and integrity
- Data quality and bias
- Third-party risk exposure
- Security vulnerabilities

Trust in AI decisions depends on trust in the components that produce them. The companion paper covers this same ground as the foundational "secure development / supply chain" layer of AI security. Governance defines what must be evaluated, and operational security carries that evaluation out once a model is running.

3. Validation: Proving Controls Work

A control that exists only on paper is an assumption. Organizations need validation that actually tests the control:

- Penetration testing
- Adversarial attack simulation
- Red teaming of AI systems
- Scenario-based stress testing

These produce evidence that the controls work, under normal conditions and hostile ones.

Skipping this step has consequences already on the record. Published research has bypassed guardrails and safety alignment that were assumed to work, using prompt-reformulation techniques that get a model to treat adversarial instructions as trusted system directives, and tokenizer manipulation that blinds a protection model before it ever fires. In one publicly documented case, an AI-powered physician assistant platform certified against SOC 2, HIPAA, ISO 27001, and GDPR was fully compromised in under two hours, with no malware, no stolen credentials, and no traditional exploit. The certifications were genuine. What none of them had done was test whether the controls held. Governance that accepts policy documentation as proof of control produces a compliance mirage.

4. Active Defense: Understanding the "State of the State"

AI systems operate in environments that keep moving, so a one-time validation goes stale. Governance requires continuous monitoring and active defense so organizations can:

- Detect anomalies in model behavior
- Identify drift or degradation
- Alert on adversarial attacks
- Respond to emerging threats in real time

Active defense and guardrails are different things. Guardrails are rules, filters, and boundary checks applied at runtime, and they answer one question: does this input or output match a known bad pattern? Active defense asks the harder, more governance-relevant question of whether the system is behaving the way it should right now. Adversaries bypass guardrails by attacking the gaps between the patterns those guardrails were built to catch, and active defense is what keeps working after a guardrail has already failed. A governance framework that treats guardrails as sufficient runtime control is certifying a static defense against an adversary who keeps moving.

That gives leaders a live view of system performance and risk exposure, the "state of the state." Informed risk decisions about AI systems depend on it, and so does any assurance about the decisions those systems make.

VI. Risk, Decisions, and Accountability

Governance exists so leaders can answer a few basic questions:

- What is an acceptable risk?
- What is not an acceptable risk?
- What actions must be taken in response to risk?
- How do I know I am at a level of proper control?

AI complicates every one of them. Its decisions can be opaque, fast, and repeated across thousands or millions of interactions, so a misaligned risk tolerance does far more damage than it used to.

So, AI governance has to make sure that:

- Risk thresholds are clearly defined
- Decision-making systems operate within those thresholds
- Deviations are detected and addressed immediately

Accountability must extend beyond individuals to include systems, models, and the processes that govern them

VII. The Rise of Agentic AI: A Governance Inflection Point

Agentic AI changes the picture again. Where traditional systems support human decision-making, agentic systems:

- Act autonomously
- Interact with other systems
- Execute multi-step processes
- Directly influence business outcomes

They now participate in the enterprise. Agentic AI enables decisions, implements them, and in some deployments holds the decision authority itself.

The operational risk is concrete. A compromised agent, operating entirely on legitimate credentials and approved business logic, becomes autonomous attack infrastructure: reading repositories, invoking APIs, moving or altering data, all without tripping a single traditional security control. Governance for agentic systems should assume that threat model from the start, well before deployment.

That creates governance problems with no clean precedent:

- **Autonomy Risk:** Systems making decisions without direct human intervention
- **Interdependency Risk:** Cascading effects across interconnected processes
- **Outcome Risk:** Direct impact on customers, operations, and financial performance

Governance therefore has to move from periodic oversight to persistent assurance, which is impossible without continuous proof of AI security.

The question is no longer: "Was this decision approved?"
It becomes: "Can we continuously ensure that all decisions, human and machine, remain aligned with our intent, risk tolerance, and obligations?"

VIII. Designing Governance for an AI-Driven Enterprise

Meeting these challenges takes a rethink of how governance is designed:

Unify Decision Frameworks:
Integrate the silos between corporate, IT, and security governance. Establish a shared model for decision-making and accountability.

Embed Governance into Systems
Governance cannot sit outside the process. It must be integrated into:

- Development pipelines
- Deployment workflows
- Runtime operations

Shift from Periodic to Continuous
Annual reviews and quarterly audits cannot keep pace. AI governance, and with it AI security, has to operate in detail and in real time.

Prioritize Evidence Over Intent
Policies define intent and committees hold meetings, but governance requires proof. Every AI security control should be measurable, testable, and demonstrable.

IX. The Future of Governance: Precision and Confidence

The goal of governance has always been the same: to let organizations operate effectively, manage risk, and meet their objectives. The level of precision required has changed.

In an AI-driven world, governance must provide:

- Granular security visibility into AI systems and their decisions
- Continuous validation of AI security controls
- Real-time insight into risk and performance

That is what decision assurance rests on.

X. Final Analysis: AI Governance Is a Competitive Advantage

Plenty of organizations will adopt AI. The ones that succeed will be those that govern it well. They will:

- Understand their AI footprint in detail
- Validate and monitor the security of their AI systems continuously
- Align decisions across business, IT, and security
- Act quickly and confidently based on real-time insight

That turns corporate governance into a strategic asset. Governance earns its place by producing better decisions and by keeping accountability and oversight attached to them.

In the age of AI, competitive advantage comes from better decisions made faster, backed by real assurance grounded in detailed AI security control.

Governance answers who decides, and how we know the decision was sound, while security answers whether the system executing that decision can be trusted to behave as intended once it is live and under attack. Neither is sufficient alone. As the companion analysis to this paper puts it: when trust has no security, AI risks everything. The reverse is equally true: when security has no governance, no one is accountable for what it protects.

Bibliography

References

1. Malcolm Harkins, “When Trust Has No Security, AI Risks Everything,” Institute for Critical Infrastructure Technology, July 2026, https://www.icitech.org/post/when-trust-has-no-security-ai-risks-everything.
2. HiddenLayer, “Prompt Puppetry: Why Every AI System Is Now at Risk,” HiddenLayer, n.d., https://www.hiddenlayer.com/insight/why-ai-systems-are-at-risk.
3. HiddenLayer, “Novel Universal Bypass for All Major LLMs,” HiddenLayer, n.d., https://www.hiddenlayer.com/research/novel-universal-bypass-for-all-major-llms.
4. HiddenLayer, “The TokenBreak Attack,” HiddenLayer, 2025, https://www.hiddenlayer.com/research/the-tokenbreak-attack.
5. National Institute of Standards and Technology, “AI Risk Management Framework,” NIST, n.d., https://www.nist.gov/itl/ai-risk-management-framework.
6. International Organization for Standardization, ISO/IEC 42001:2023: Information Technology — Artificial Intelligence — Management System (Geneva: International Organization for Standardization, 2023), https://www.iso.org/standard/42001.
7. Organisation for Economic Co-operation and Development, “AI Principles,” OECD, n.d., https://www.oecd.org/en/topics/sub-issues/ai-principles.html.

Author

Malcolm Harkins

Chief Security & Trust Officer,
Hidden Layer

Malcolm Harkins is the Chief Security and Trust Officer at HiddenLayer. He has over two decades of experience in information security leadership roles at top technology companies, including Intel, Cylance, and others. He's written multiple books on risk management, information security, and IT and earned awards from the RSAC, ISC2, Computerworld, and in May of 2026 was inducted in the CSO Hall of Fame.

Harkins has testified before the Federal Trade Commission and the U.S. Senate Committee on Commerce, Science, and Transportation. Harkins is a Fellow with the Institute for Critical Infrastructure Technology, a non-partisan think tank providing cybersecurity expertise to the House of Representatives, Senate, and various federal agencies. He holds a bachelor's degree in economics from the University of California at Irvine and an MBA in finance and accounting from the University of California at Davis. Harkins also taught at UCLA's Anderson School of Management and Susquehanna University.