O'REILLY®
Report

Securing AI Systems

A Comprehensive Framework for Enterprise Defense

Pamela K. Isom

cyberhaven

Unified AI & Data Security Platform

- Secure Agentic and Human Workflows automatically
- Replace four point solutions with one for comprehensive data security
- See where your data goes, even after it leaves with end-to-end Data Lineage

Learn more at cyberhaven.com

Securing AI Systems

A Comprehensive Framework for Enterprise Defense

Pamela K. Isom

O'REILLY®

Securing AI Systems
by Pamela Isom

Published by O'Reilly Media, Inc., 141 Stony Circle, Suite 195, Santa Rosa, CA 95401.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*https://oreilly.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Nicole Butterfield
Development Editor: Gary O'Brien
Production Editor: Aleeya Rahman
Copyeditor: Charles Roumeliotis

Cover Designer: Susan Brown
Interior Designer: David Futato
Interior Illustrator: Kate Dullea

June 2026: First Edition

Revision History for the First Edition
2026-06-16: First Release

This work is part of a collaboration between O'Reilly and Cyberhaven. See our statement of editorial independence.

979-8-341-67383-0

[LSI]

Table of Contents

CHAPTER 1
The AI Security Imperative

Artificial intelligence (AI) is rapidly becoming part of everyday business operations. Enterprise platforms now incorporate AI capabilities that assist with analysis, communication, decision support, and workflow coordination. In many organizations, the digital workforce now includes not only employees but also AI systems and autonomous software agents that interact across enterprise systems.

Adoption is accelerating quickly. According to the Stanford Institute for Human-Centered Artificial Intelligence (HAI) AI Index Report 2025, and illustrated in Figure 1-1, 78% of organizations now use AI in at least one business function, up from 55% the previous year.[1]

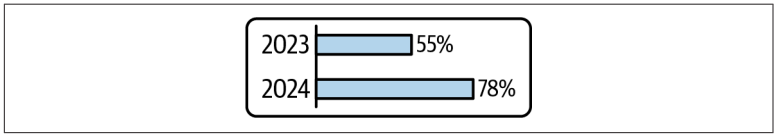

Figure 1-1. Rapid growth of enterprise AI adoption

As AI becomes embedded in core workflows, a new security imperative emerges. Organizations must ensure that AI-driven processes operate within defined governance boundaries, that automated actions remain accountable, and that sensitive information is not unintentionally exposed through AI-enabled interactions.

1 Nestor Maslej et al., *The AI Index 2025 Annual Report* (Stanford, CA: AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 2025), *https://hai.stanford.edu/ai-index/2025-ai-index-report*.

The following scenario illustrates how these challenges can arise in everyday enterprise environments.

Understanding the AI Security Challenge

A leadership team is preparing for an upcoming executive review. To accelerate preparation, a project manager asks the organization's AI assistant to compile a briefing summarizing recent performance reports, operational updates, and emerging risks.

The assistant is connected to several enterprise systems and is authorized to retrieve internal documents, analyze operational data, and summarize key findings. As part of this process, the AI coordinates with other specialized agents responsible for retrieving data, reviewing documentation, and preparing workflow updates across enterprise systems.

Within seconds, the system produces a clear report highlighting performance trends, operational concerns, and recommended next steps. It also proposes notifying regional teams about potential delays and scheduling follow-up tasks in the organization's workflow platform. The briefing appears useful and well written, and the manager includes it in the presentation.

Later, a question arises.

Several insights in the briefing appear to reference internal research materials and operational reports that were never intended to be widely shared. Some recommendations appear to reflect patterns identified across multiple internal sources. Security teams are asked to review the interaction.

Investigators begin asking questions:

- What information did the AI system access while preparing the report?
- Did it combine internal and external data sources?
- Did any proprietary information leave the organization during the interaction?
- Did the system rely on internal records that should not have been used in this context?

- Did the AI assistant trigger additional automated actions based on its recommendations?

In many organizations, the answers are difficult to determine.

Modern AI systems do more than retrieve information. They interpret data, generate insights, orchestrate actions, and increasingly operate through networks of specialized agents that interact across enterprise systems. These capabilities can dramatically improve productivity and decision support, but they also introduce new questions about visibility, accountability, and control.

For executives and security leaders, the challenge is not simply whether AI systems are accurate or useful. The challenge is whether organizations can maintain visibility into how information is accessed, interpreted, and used as AI systems increasingly participate in operational workflows.

Addressing the Challenge

The chapters that follow examine how organizations can address this challenge.

Chapter 2, "An AI Security Operating Model Supported by Established Frameworks", introduces an operating model for governing and securing AI systems across the enterprise. It outlines five reinforcing pillars that help organizations understand AI usage, protect sensitive data, define acceptable policies, enforce safeguards during use, and continuously monitor system behavior.

Chapter 3, "Moving Beyond Traditional Data Loss Prevention", examines why traditional data protection tools cannot fully secure AI workflows and explains what modern AI data security must provide to restore visibility and control.

Chapter 4, "Implementing AI Security in Practice", provides a practical roadmap for implementing an AI security program. It translates governance principles into operational steps organizations can take to establish visibility, apply controls, and sustain oversight as AI adoption expands.

Together, these chapters present a structured approach to securing AI-enabled environments while enabling organizations to benefit from the speed and innovation AI provides.

CHAPTER 2

An AI Security Operating Model Supported by Established Frameworks

Artificial intelligence changes how organizations experience risk. Traditional systems execute predefined instructions, but AI systems generate outcomes through inference and interaction. As organizations adopt these capabilities, they must secure not only systems but also decisions and influence. During normal use, employees and automated workflows send fragments of data to external models and embedded services. Individually these fragments appear harmless, yet AI systems recombine them and affect business outcomes. Security therefore depends on where data is stored, how it moves, how systems interpret it, and how results drive action.

Organizations cannot rely solely on traditional cybersecurity practices. Leaders must coordinate governance, operational oversight, and technical safeguards across the AI lifecycle—sourcing, training, release, and operation. The lifecycle shows where risk forms; however, you need a repeatable method to control it. This chapter introduces an AI Security Operating Model that organizes protection into five reinforcing pillars: AI usage and shadow AI discovery, understanding data and lineage, defining AI-aware policies, enforcing controls at the point of use so that protections are applied while a user is interacting with AI, and continuously monitoring and improving performance and usage of AI models and governance programs across an organization.

The following sections first describe the operating model and then connect it to governance responsibilities, lifecycle risk domains, and established security frameworks. Ad hoc experimentation leaves organizations exposed because tools evolve faster than governance does. Mature enterprises secure AI as part of the systems, data, and decisions it influences, with clear ownership and oversight guiding how it is deployed and used. When leadership defines accountability and operational teams enforce controls, organizations can expand AI use while maintaining confidence in outputs and protecting sensitive information. This shift establishes the foundation for a practical operating model.

The Five Pillars of an AI Security Program

AI risks require more than isolated safeguards; they require an operating model. Organizations need a structure that assigns responsibility, governs data interaction, and sustains oversight as systems evolve. This framework organizes AI security into five reinforcing pillars that translate enterprise AI risk into operational responsibilities. Rather than treating security as a one-time review or purely technical activity, the model embeds protection into how the organization discovers, governs, uses, and continuously evaluates AI.

As shown in Figure 2-1, the framework begins with AI Usage and Shadow AI Discovery, which establishes visibility into AI assets and unsanctioned usage. It continues with the second pillar, Understand Your Sensitive Data and Its Lineage, which keeps information interacting with AI controlled and trustworthy. The third pillar, AI-Aware Security Policies, defines acceptable use and decision boundaries for probabilistic systems. Organizations then Enforce Controls at the Point of Use (the fourth pillar), applying safeguards such as access restrictions, real-time validation, and intervention mechanisms. Finally, the fifth pillar, Monitor, Investigate, and Continuous Improvement, sustains the program by using behavioral changes and anomalies to drive ongoing risk management. Together, these pillars move organizations from reactive protection to a repeatable AI security operating model.

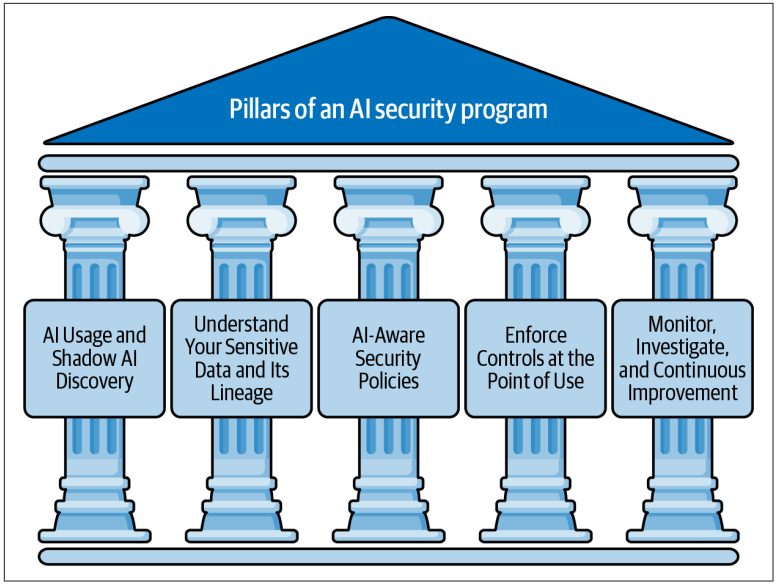

Figure 2-1. Five pillars of an AI security program

NOTE

Scope of AI Security

AI security extends beyond standalone AI tools. Predictive and generative capabilities now exist across everyday business platforms. It requires securing business outcomes, not applications, which means understanding how automation influences decisions across the enterprise.

To understand what these pillars govern, let's first examine the organizational responsibilities and lifecycle risk domains that shape AI exposure:

- **Governance and oversight responsibilities**

  Risk ownership and accountability
  : Leadership assigns ownership of AI outcomes and defines authority to approve, restrict, or terminate operation.

  Policy and standards
  : The organization establishes acceptable use, safety thresholds, and security requirements aligned with legal and operational obligations.

Independent assurance
: Independent reviewers conduct audits, red-team exercises, and oversight separate from system developers and operators.

Incident response authority
: Leadership defines escalation paths and decision rights when models behave unexpectedly or unsafely.

- **Key lifecycle risk domains**

  Supply chain assurance
  : Verify the origin, integrity, and dependencies of AI models, components, and training data.

  Data governance and management
  : Ensure data remains authoritative, appropriate for its intended use, properly handled throughout its lifecycle, and protected to preserve privacy, regulatory compliance, and decision integrity.

  Secure development
  : Design models with embedded security, privacy, and resilience requirements, including safeguards against misuse and adversarial manipulation.

  Release validation and adversarial testing
  : Evaluate systems for quality and safety prior to approval and deployment by simulating realistic misuse and manipulation to identify unsafe behaviors and manipulation techniques. For example, attackers may embed hidden instructions within web pages that an AI system later retrieves, influencing downstream recommendations or decisions without compromising the model itself.

  Continuous monitoring and operational oversight
  : Monitor deployed models and autonomous systems for drift, behavioral changes, and policy violations and trigger review, rollback, retraining, or containment when thresholds are exceeded.

To translate these governance responsibilities and lifecycle risks into actionable practice, begin by establishing visibility into how AI is actually used across the enterprise. Before policies can be enforced,

safeguards applied, or oversight sustained, leaders need a clear understanding of where AI systems exist, how they are accessed, and where unsanctioned or unmonitored use may introduce hidden exposure. This is why the operating model commences with Pillar 1: AI Usage and Shadow AI Discovery.

Pillar 1: AI Usage and Shadow AI Discovery

AI security begins with visibility. Organizations cannot secure systems they do not know exist, and AI adoption often begins informally through employee experimentation, external tools, and embedded AI features within existing software. This creates *shadow AI*—unsanctioned usage that introduces data exposure, decision risk, and untracked dependencies. In most cases, it arises not from negligence but from pressure to improve speed, efficiency, and quality faster than governance processes can respond.

This pillar establishes continuous discovery of AI usage across the enterprise. Organizations should maintain an inventory of sanctioned and unsanctioned AI tools, including public generative services, embedded software features, APIs, and experimental pilots. Examine roles, departments, devices, and access methods to understand who uses AI, how it is used, and where it influences business processes. Because AI capabilities evolve rapidly, discovery is not a one-time inventory but an ongoing activity that identifies newly introduced tools, updated features, and emerging usage patterns. Extend discovery beyond employee-initiated tools to include AI embedded within enterprise software and third-party services. Many vendors now incorporate generative or predictive capabilities that process organizational data without explicit user awareness. Teams should therefore maintain visibility into the AI adopted and the AI operating within the supply chain. This approach aligns with established risk frameworks that treat asset and supplier awareness as prerequisites to managing risk.

Visibility enables risk-based enablement rather than restriction. Understanding usage patterns provides approved environments for experimentation, such as controlled testing sandboxes, while preventing sensitive data exposure to untrusted services. Training reinforces this visibility by helping employees recognize AI-specific risks such as prompt manipulation and AI-assisted phishing. The outcome is operational awareness: leadership maintains a reliable view of AI adoption, and security decisions rely on observed

behavior rather than assumptions. Without discovery, organizations operate blindly.

Pillar 2: Understand Your Sensitive Data and Its Lineage

Today, sensitive data extends far beyond traditional identifiers such as Social Security numbers or payment details. Stewardship requires defining what requires protection within our operations, including proprietary intellectual property, internal decision data, operational records, and even outputs generated by AI systems themselves. Without a clear definition, safeguards cannot function. In AI environments, security and data governance become inseparable. Traditional software behaves according to code; AI behaves according to data. Protecting the system therefore means controlling what the model receives, what it learns, and what it can later reveal.

AI systems routinely bridge environments. They pull information from software as a service (SaaS) platforms, local endpoints (operator consoles, engineering workstations, monitoring dashboards), and cloud repositories, then process it within external or embedded models. During normal use, fragments of organizational data may leave governed systems and enter third-party services. Risk therefore depends not only on storage location but on data movement and interpretation. This is managed through *data lineage*. Lineage shows where data originated, how it was transformed, and where it influenced decisions. When a model produces incorrect, unsafe, or biased output, lineage helps us determine whether the issue arose from the source data, the processing context, or the model's behavior. AI creates meaning by combining information. An internal troubleshooting guide, a system hostname, and an employee contact list may each appear low sensitivity individually. When combined in a single interaction, the model can infer system architecture and operational responsibilities. Static labels would allow the exchange; contextual classification prevents it.

AI systems retain history and synthesize relationships across prompts, documents, and sessions. Information that appears harmless in isolation becomes sensitive through aggregation. This pillar requires evaluation of what data is and what it reveals when combined. AI-assisted classification supports this protection by:

- Identifying anomalous access patterns and data movement
- Distinguishing normal processing from AI prompt submission

- Preventing confidential information from entering external models

Use lineage and contextual classification to inform governance decisions about acceptable use and required safeguards. The outcome of this pillar is traceability and control: understand what data influences AI decisions and where exposure can occur. Without this visibility, you can't verify integrity nor contain exposure.

Pillar 3: AI-Aware Security Policies

With visibility into AI usage and understanding of the data it touches, you can define acceptable risk. Pillar 3 establishes decision authority by translating awareness into enforceable rules governing how AI may influence the organization.

Traditional security programs rely on binary allow-or-block decisions. In AI environments, blanket prohibitions drive activity into unsanctioned channels and reduce visibility. Instead, define permitted, restricted, and prohibited uses based on risk.

Organizations may restrict certain tools, but restriction without alternatives creates workarounds. When employees still need capability, they turn to personal accounts or external services, increasing exposure. Therefore, pair limits with approved pathways. For example, consider allowing a generative AI tool for public marketing content while prohibiting its use as an AI agent that will handle proprietary code and regulated data. This approach protects the enterprise without undermining productivity. Further, agentic AI tools combine generative AI with other autonomous capabilities to carry out tasks and introduce additional layers of complexity for security policy development. It is important, and a teachable moment, to convey to employees that AI use cases scale beyond speed and to provide concrete examples of data sensitivity risks.

We evaluate risk across four dimensions:

Data sensitivity
: What information the system processes

Tool characteristics
: Whether the model is public, external, or enterprise-controlled

User role and purpose
: Why the AI is used and by whom

Decision impact
: Consequences if the output is incorrect

Not all AI functions carry equal risk. A spelling suggestion presents minimal exposure, while automated financial, medical, or security decisions require strict oversight. Teams should scale requirements proportionally and define when testing, approval, or human review is required.

Governance operates through multidisciplinary review. Security, business, data, and operational stakeholders jointly evaluate new capabilities before approval and define acceptable-use conditions and escalation authority. Leadership sets risk tolerance and approval authority; operational teams configure and enforce controls. Separating governance from management keeps policy intentional rather than improvised through tooling.

Because AI behavior emerges during use rather than static code, approval evaluates behavior, not only specifications. Instead of asking "Is this application allowed?" teams should ask:

- What can the system learn?
- What can it infer?
- What can it retain?
- What actions can it influence?
- Where might it be considered?

This shifts the approval from software vetting to behavioral risk assessment.

The outcome of this pillar is organizational clarity. Employees understand how AI may be used, leadership defines acceptable risk, and controls enforce decisions consistently across the enterprise.

Pillar 4: Enforce Controls at the Point of Use

Policies reduce risk only when applied where work occurs. In this pillar, safeguards operate during interaction rather than after an incident. A credible practice is to intervene while decisions form, not after they are executed. AI systems operate faster than traditional review processes, so enforcement spans browsers, endpoints, operational applications, and SaaS platforms where people interact

with AI. When a user attempts to submit sensitive data to an external model, controls act before exposure occurs.

Enforcement balances prevention with guidance. High-risk actions are blocked, while lower-risk situations trigger real-time coaching. Users receive warnings when activity approaches policy boundaries and are directed to approved alternatives. This reduces accidental exposure without disrupting productivity. A common exposure path involves employees using personal AI accounts for business tasks. Controls distinguish enterprise environments from personal services and apply different protections to each. Data remains within approved channels instead of drifting into unmanaged systems. Policy therefore follows the user wherever AI interaction occurs.

Operational Example: Controlled Decision Authority in Financial Workflows

Consider an agentic AI assistant that drafts payment approvals and flags financial anomalies. Without protections applied while a user is interacting with AI, a manipulated prompt or poisoned data source could lead the agent to recommend releasing funds to an attacker. The agent may prepare recommendations but cannot authorize transactions, change payment destinations, or bypass approval thresholds. A human reviewer approves the action, and the agent cannot silently alter the outcome. The primary risk is not incorrect output; it is uncontrolled influence. When AI recommendations directly trigger business actions, the organization loses decision authority. Point-of-use controls constrain what the AI can do during operation. The objectives are accuracy and controlled behavior, preserving decision integrity even under manipulation, corrupted data, or misplaced trust.

The outcome of this pillar is real-time enforcement. Organizational rules shape behavior during interaction and prevent unsafe actions before they occur. Because not every scenario can be predicted, observed behavior feeds refinement of safeguards, leading to sustained monitoring, investigation, and improvement.

Pillar 5: Monitor, Investigate, and Continuous Improvement

In this pillar, the program moves from enforcement to assurance. AI systems combine deterministic software behavior with probabilistic outcomes, so security cannot end at deployment. A system that behaves correctly today may behave differently tomorrow due to changes in system behavior over time, data changes, environmental shifts, or adversarial influence. Governance therefore operates as a continuous cycle rather than a linear process.

Security becomes an operational signal instead of a periodic audit. Monitoring behavior, not just system uptime, evaluates inputs and outputs and detects deviations from expected performance. Anomaly detection provides early warning of manipulation attempts, data leakage, model degradation, or misuse.

Monitoring alone does not create trust, explanation does. When anomalies occur, we reconstruct the interaction: which prompts were submitted, what data entered the system, and how the model responded.

This visibility provides traceability and explainability, allowing reviewers to determine whether the issue resulted from malicious activity, model failure, or human error. Findings feed formal incident and risk management processes, so responses remain consistent and defensible. Automated correlation across logs, prompts, lineage data, and system activity accelerates investigation while preserving evidence for compliance and impact assessment.

The objective is detection and improvement. Each event strengthens the system through:

Policy refinement
: Update AI-use boundaries based on observed behavior

Control adjustment
: Tune safeguards and thresholds using real-world usage patterns

Workforce learning
: Shift training to operational awareness instead of periodic compliance

Review events across functions. When operational, security, legal, and business teams investigate separately, they correct symptoms rather than causes. Cross-functional review connects technical findings to decision impact and governance accountability. As adoption expands, risk tolerance evolves. New use cases, increased reliance on automation, and observed behavior reshape acceptable risk. Continuous monitoring therefore informs both technical safeguards and leadership decisions about where automation is appropriate and where human oversight must remain.

By treating monitoring and investigation as learning mechanisms rather than purely defensive measures, the organization moves from reactive protection to sustained assurance, maintaining confidence in system behavior, protecting decision integrity, and enabling responsible innovation over time.

Alignment with the NIST Cybersecurity Framework and AI Profile

Stakeholder interviews consistently showed the need for established frameworks to determine which AI capabilities belong in the enterprise and how their risk should be governed over time. The five pillars in this chapter provide the operating model, while recognized standards such as the NIST Cybersecurity Framework (CSF) 2.0 and the Cybersecurity Framework Profile for Artificial Intelligence (NIST IR 8596 [draft]) provide the reference structure for evaluating completeness and maturity.

The NIST framework organizes cybersecurity risk management across six functions: Govern, Identify, Protect, Detect, Respond, and Recover. These functions define what an organization must achieve: accountability, awareness of exposure, safeguards, monitoring, incident handling, and restoration of trust.

The five pillars operate at a different level. Rather than introducing additional requirements, they translate these functions into repeatable organizational behavior across AI discovery, data interaction, decision authority, operational enforcement, and continuous assurance. In practice, organizations operate through the pillars and assess effectiveness through NIST. Table 2-1 shows how the five pillars of an AI security program map to the NIST cybersecurity functions and their high-level purposes.

The distinction is practical: the framework defines required capabilities, while the operating model defines how those capabilities function day to day. For example, the NIST Govern function establishes accountability and risk tolerance; within the pillar model, this translates into decision authority, ownership of AI outcomes, and defined human oversight. Likewise, Detect, Respond, and Recover become continuous monitoring, investigation, and corrective action with proactive reviews and applied lessons learned, keeping automated behavior aligned with business intent.

Table 2-1. Translating framework expectations into operational accountability: mapping formal cybersecurity functions to the organizational responsibilities required to govern and secure AI use

Pillars	Corresponding NIST functions	NIST purpose
AI Usage and Shadow AI Discovery	Identify	Understand AI assets and exposure
Understand Data and Lineage	Identify, Protect	Protect information integrity
AI-Aware Policies	Govern, Protect	Define acceptable behavior
Enforce at Point of Use	Protect, Detect, Respond, Govern	Prevent unsafe actions in real time
Monitor and Improve	Detect, Respond, Recover, Govern	Maintain trust over time

The AI profile further emphasizes risks unique to probabilistic systems (systems that generate answers rather than follow fixed rules), including model manipulation, data lineage exposure, and behavioral drift—reinforcing that AI assurance cannot rely on one-time approval but requires ongoing evaluation. Together, the approaches are complementary. NIST measures whether protection is complete; the five pillars make protection operational.

This distinction exposes a practical gap. Organizations can define governance, assign responsibility, and document acceptable use, yet still lack the ability to apply those decisions during real AI interaction. Most legacy security tools were designed to protect files and systems, not how AI interprets, combines, and influences information. As a result, the challenge shifts from defining policy to enforcing intent. Securing AI requires controls that operate on data meaning, context, and decision influence rather than location alone.

The next chapter examines why traditional data protection approaches fall short in AI environments and what modern data-centric security must provide to operationalize this model in practice.

CHAPTER 3

Moving Beyond Traditional Data Loss Prevention

Organizations evaluating AI security often begin with a reasonable question: *can existing data loss prevention (DLP) tools solve this problem?* For most enterprises, the answer is no. Traditional DLP technologies were designed for a different era of computing, one in which sensitive information resided in identifiable files, structured databases, and well-defined network boundaries. AI fundamentally alters how information moves, how it is interpreted, and how risk materializes. Securing AI workflows therefore requires moving beyond traditional DLP. Organizations must evolve from monitoring document transfers to governing what and how information combines, influences AI systems, and shapes automated decisions.

Traditional DLP Assumptions

Legacy DLP tools focus on detecting full document transfers or structured data leaving controlled environments. These systems typically rely on predefined classification rules, for example, triggering an alert if a document contains a Social Security or address pattern. This approach assumes that sensitive data exists in identifiable files or structured records that can be monitored as they move across systems.

In practice, this model worked reasonably well for traditional workflows in which employees emailed documents, transferred files, or

accessed structured databases. However, AI workflows rarely follow this pattern.

How AI Workflows Break Traditional DLP

As discussed in Chapter 2, modern AI workflows often involve employees sharing small portions of information through conversational interactions rather than transferring full documents. Industry analysis reflects this shift: the most common types of data employees input into AI tools across industries are source code (8.3%), research materials (10.7%), and human resources (6.2%).[1] Overall, 39.7% of all human interactions with AI tools involve sensitive data. These inputs are typically copied into conversational workflows as users request research or generated content.

AI security therefore shifts risk from document transfer to contextual recombination, where multiple aspects of information combine to reveal sensitive insights. This creates a growing risk of accidental intellectual property exposure, where proprietary knowledge is shared outside controlled environments without malicious intent.

Encryption and traditional DLP remain essential safeguards for protecting data at rest and in transit, but AI workflows introduce risk *after information is decrypted for legitimate use*.

Traditional DLP assumes files have a creator, a location, and a system of record. Ownership is traceable. AI systems, however, generate derivative content—summaries, recommendations, rewritten documents, and code suggestions.

Consider a simple example. An AI assistant summarizes five internal documents into a new executive brief. The resulting output contains proprietary insight, yet it may have no original classification label and ambiguous ownership.

Several questions should immediately arise:

- Who owns the derivative output?
- What classification applies?
- Can it be shared externally?

1 Cyberhaven Labs. *2026 AI Adoption and Risk Report*, *https://www.cyberhaven.com/resources/report/ai-adoption-risk-report-2026*.

Traditional DLP does not track derivation. By default, it cannot determine how information flowed into the output or whether sensitive insight was embedded along the way.

The question quickly becomes one of governance: *who owns the underlying data, who controls the resulting models, and how the information they generate is ultimately used.* Without lineage visibility into how information flows, organizations cannot answer these questions or demonstrate that sensitive information was handled appropriately.

NOTE

AI security cannot rely solely on blocking tools or restricting use. Organizations must shift from file-based protection to data-centric governance that controls how information is interpreted, recombined, and used in automated decisions.

This move toward data-centric governance aligns with the AI security operating model introduced in Chapter 2, where we outlined the five pillars, in which organizations establish visibility into AI usage, govern the data that interacts with the models, define acceptable use, enforce safeguards at the point of interaction, and continuously monitor outcomes.

From File-Centric to Data-Centric Security

In AI interactions, sensitive exposure may not involve a recognizable file at all. A model may receive a paragraph of internal strategy, a configuration snippet, and a support log entry across separate interactions. Individually, this data may appear low risk. Collectively, they can reveal operational insight, such as internal system architecture, configuration weaknesses, or decision thresholds.

- A file-centric approach cannot detect this exposure because no single document contains the risk.
- A data-centric approach instead follows the information itself, regardless of format, channel, or destination.

Table 3-1 highlights key structural differences between traditional DLP and AI data security. Legacy controls monitor transfer, but AI risk emerges through interpretation and recombination.

Table 3-1. Structural differences between traditional DLP and AI data security

Traditional DLP	AI data security
Protects documents	Protects how data is interpreted
Detects file transfer	Detects contextual recombination
Uses static classification	Uses context and data lineage
Assumes clear file ownership	Manages derivative outputs
Monitors specific channels	Tracks influence across workflows
Protects data at rest and in motion	Protects how data influences automated decisions

What Organizations Must Do Differently

This shift requires organizations to evolve their security posture. Organizations must *reframe risk from "data leaving systems" to "data influencing decisions."* Exposure may occur without a file transfer event. Governance must address how information shapes automated outcomes:

Require traceability of data influence
: Security leaders should evaluate whether their platforms can answer several essential questions:

  - Where did this information originate?
  - How did it enter the AI interaction?
  - How did it influence the resulting output?
  - What policies and controls governed its use?
  - Who approved the resulting action?

  If you cannot answer these questions, DLP is incomplete.

Apply consistent policy enforcement across AI and traditional systems
: AI interactions cannot be governed separately from broader data protection programs. Security must extend across endpoints, browsers, SaaS platforms, and AI systems.

Align oversight with business risk tolerance
: Not every AI interaction carries equal consequence. High-impact use cases such as financial approvals, regulatory reporting, or operational control require proportionate safeguards and monitoring.

As organizations adopt AI agents and agentic workflows that orchestrate and execute tasks across systems, compromised or manipulated inputs may influence automated actions. Governance must therefore preserve decision authority and ensure automated systems operate within defined boundaries. If, for instance, customer travel information is exposed, how soon should organizations be notified, and what are the decision rights for both humans and the AI? What are the strategies for testing for DLP using derivative scenarios? This subject will be discussed more in Chapter 4.

Measure governance maturity, not tool presence
: The relevant question is not whether an organization has DLP. It is whether the organization can trace, govern, and constrain how data influences automated decisions.

This shift also expands the role of the chief information security officer. In AI environments, the responsibility extends beyond protecting systems and data stores to protecting the influence of data on automated decisions. As AI systems participate in analysis, recommendations, and operational workflows, security leaders must ensure that data remains trustworthy, decisions remain accountable, and automated behavior operates within defined governance boundaries. AI security therefore represents an evolution of the security mission, not a departure from it.

NOTE

Data centric governance is not optional. Without it, organizations risk maintaining compliance optics while losing operational control.

The Context and History Problem

AI systems do not process information in isolation. They retain conversational context and formulate outputs based on cumulative interaction history. Traditional DLP tools often lack visibility into:

- What data was previously shared
- How multiple interactions combine meaning
- Whether outputs contain sensitive derivatives
- How data flows from source systems to AI tools and downstream processes

Without historical context, organizations cannot determine whether exposure has occurred, let alone reconstruct how it happened. Leadership may be unable to explain how data influenced a decision or trace the origin of problematic output.

Addressing this challenge requires more than improved alerting. Modern AI security programs must provide the ability to:

Trace how information moves from source systems into AI interactions
: Identify the originating systems, detect when content is submitted to AI tools, and trace where derivative outputs are later stored, shared, or acted upon.

Correlate fragments across sessions and applications
: Exposure rarely occurs in a single action. Security controls must recognize when multiple pieces of information combine across conversations, systems, or time to reveal sensitive insight.

Apply policy to derivative outputs, not only original inputs
: Extend protection to AI-generated content that contains embedded proprietary or regulated information.

Preserve accountability for automated recommendations
: Automated systems should operate within defined authority boundaries, require human oversight for high-impact actions, and maintain auditable records of what data influenced the outcome.

Modern AI data security platforms increasingly integrate capabilities traditionally separated across DLP, insider risk monitoring, data security posture management (DSPM), and AI interaction oversight. This unified visibility allows organizations to track how information moves and influences automated outcomes across systems rather than monitoring each risk domain independently.

Summary

AI operates in a context-driven and inference-based environment. Risk no longer emerges solely from document transfer, but from how information is recombined, interpreted, and embedded in automated decisions. Traditional DLP can detect file movement, but it cannot reliably trace how data influences outcomes across conversational prompts, derivative outputs, and agentic workflows.

Without contextual visibility and data lineage, organizations cannot explain how sensitive and even not-so-sensitive information shaped an AI-generated outcome or demonstrate that it was handled appropriately.

AI security therefore requires a structural shift from file-centric monitoring to data-centric governance that preserves traceability, enforces policy across interaction channels, and maintains accountability for automated outcomes.

Having defined the structural gap, the next question becomes practical: how should organizations implement an AI security program that restores control while enabling innovation?

CHAPTER 4

Implementing AI Security in Practice

As organizations integrate AI-enabled capabilities into business operations, the focus shifts from understanding risk to governing how AI systems operate in practice. Security and governance are not separate efforts. They must be embedded directly into AI adoption decisions so that innovation and risk management evolve together.

A central challenge is maintaining visibility across the enterprise. Traditional inventories focus on systems and applications, but AI introduces additional layers including data, models, agents, integrations, and workflows that are more dynamic and less transparent. Without this visibility, governance cannot be consistently enforced as adoption scales.

This challenge is amplified by how AI is already improving day-to-day operations. Teams are using AI to synthesize insights across large volumes of interaction data, including customer conversations, operational records, and internal communications. What once required significant manual effort can now be performed in seconds, enabling faster understanding of trends, risks, and opportunities. In many cases, users begin from an advanced draft or pre-analyzed state rather than starting from scratch.

These capabilities are working. They increase speed, expand analytical reach, and improve productivity across functions. However, they also accelerate the ways information is combined, interpreted, and acted upon, requiring governance to operate at the same pace as AI-enabled workflows.

Implementing AI security is not a single action, but a set of coordinated decisions that begin with selecting AI solutions aligned with organizational risk tolerance that ultimately define acceptable use, data sensitivity, and control requirements. These early choices shape how data is handled, how systems are governed, and how effectively controls can be applied.

Organizations must also explicitly address inference risk. AI systems process data—and generate new insights, relationships, and attributes that may themselves be sensitive. These inferred outputs can expose information that was never directly collected, making risk less visible and harder to control within traditional data governance approaches.

In practice, business leaders and practitioners must move from defining where AI can operate to governing how data is accessed and used, how system behavior is controlled, how outcomes are validated, and how oversight is sustained. The sections that follow reflect this progression and provide a path for translating governance into operational reality. This progression reflects a structured approach to governing AI across its lifecycle, one defined by the operating model.

Connecting the Operating Model to Implementation

The operating model introduced in Chapter 2 provides the foundation for governing and securing AI systems. It defines the core pillars required to establish visibility, enforce control, and sustain oversight.

In practice, these pillars are realized through operational decisions embedded in how AI is selected, deployed, accessed, and monitored. Governance is not abstract; it is executed at the point of use and sustained through continuous oversight. The sections that follow translate each pillar into concrete actions, providing a practical roadmap for securing AI systems in real-world environments:

- "Aligning Leadership and Accountability for AI Security" supports the consistent application of Pillar 3: AI-Aware Security Policies and organizational accountability for how AI systems are governed.
- "Where AI Is Allowed to Operate" supports Pillar 1: AI Usage and Shadow AI Discovery, enabling visibility into both approved and unapproved AI activity across the enterprise.
- "What AI Systems Are Allowed to Access and Do" reinforces both Pillar 2: Understand Your Sensitive Data and Its Lineage and Pillar 4: Enforce Controls at the Point of Use, ensuring that data is appropriately managed and that system interactions remain within defined boundaries.
- "How Organizations Validate AI-Driven Actions" reinforces Pillar 3: AI-Aware Security Policies, ensuring acceptable use and accountability in how AI influences decisions and outcomes.
- "Managing Agentic AI in Operational Workflows" reinforces Pillar 4: Enforce Controls at the Point of Use, helping maintain control over automated behaviors and system interactions in dynamic environments.
- "Managing External Access and Workforce Expansion" extends both Pillar 3: AI-Aware Security Policies and Pillar 4: Enforce Controls at the Point of Use, ensuring that access and use align with governance expectations across internal and external participants.
- "Red Teams: Testing External Access Through AI Workflows" enables Pillar 5: Monitor, Investigate, and Continuous Improvement, validating whether controls are effective under real-world conditions.
- "Sustaining Oversight as AI Scales" enables Pillar 5: Monitor, Investigate, and Continuous Improvement, ensuring that risks are continuously identified, assessed, and addressed as AI adoption expands.

Together, these sections translate the operating model into practical actions that align directly with how AI systems are implemented, governed, and sustained in real-world environments.

Aligning Leadership and Accountability for AI Security

As AI becomes more embedded across operations, defining responsibility for governing and securing these capabilities becomes essential. The chief AI officer typically drives strategy, adoption, and business value, while the chief information security officer maintains responsibility for enterprise security and risk management. AI security spans both domains. Rather than creating separate or competing structures, a shared accountability model is required. In many cases, AI security is implemented as a specialized capability within the security organization, with strong alignment to AI leadership to ensure that controls are embedded directly into operational workflows.

Effective alignment depends on a shared understanding of organizational risk tolerance. Risk appetite defines how AI may be used, what data can be exposed, and what level of autonomy systems are permitted to operate with. Overly restrictive policies can limit adoption and drive usage into unmonitored environments. At the same time, insufficient controls can result in inconsistent decision making and unmanaged risk. Leadership must therefore establish clear boundaries that balance enablement with control, ensuring that governance reflects both operational needs and acceptable levels of risk.

In many cases, AI adoption does not introduce entirely new regulatory obligations but extends existing ones. Established frameworks governing intellectual property, data protection, financial controls, and industry-specific requirements already apply to many AI use cases.

The challenge is not operating in an unregulated environment but determining how these existing obligations apply to new forms of data use, automated decision making, and AI-generated outputs. Organizations must therefore translate established requirements into AI-enabled contexts rather than assume that AI operates outside current governance structures. This approach enables consistent enforcement of security standards while ensuring that capabilities are designed and deployed with governance in mind. It also reinforces that AI security is not owned by a single function, but requires coordination across leadership, technical teams, and business stakeholders.

Where AI Is Allowed to Operate

As AI capabilities expand across enterprise platforms, determining where these systems are permitted to operate becomes essential.

In many environments, adoption occurs organically. Tools are embedded into enterprise applications, introduced through new features, or used informally by employees. Without clear boundaries, these systems may interact with sensitive data, operational workflows, or decision processes in ways that exceed expectations.

Usage often extends beyond formally approved tools. Embedded capabilities within enterprise platforms and informal use across teams can create significant visibility gaps. Addressing this requires defining approved use and establishing mechanisms to understand how AI is being used across the organization.

Two Approaches to AI Adoption

Organizations often operate between two competing approaches to AI adoption:

- Restriction-oriented approach
  - — Limits AI use until risks are fully understood
  - — Prioritizes preventing exposure and misuse
- Enablement-oriented approach
  - — Encourages AI use to improve productivity and insight
  - — Expands use across business functions

In practice, a hybrid model is most effective because overly restrictive approaches can drive AI use into unmonitored environments, while unrestricted use can introduce unmanaged risk. Effective AI security requires *controlled enablement* within visible, governed, and secure environments.

Key considerations include:

- Where use is appropriate and where it is restricted
- Which business processes can incorporate it
- How usage aligns with risk tolerance

This includes distinguishing between approved enterprise capabilities and informal or unvetted use across business functions. Establishing these boundaries ensures that AI operates within intentional and controlled environments.

What AI Systems Are Allowed to Access and Do

Once usage is defined, the next step is determining what these systems are allowed to access, how sensitive information is handled, and what actions they are permitted to take.

AI systems may retrieve data from multiple sources, combine information across contexts, and generate outputs that influence decisions or trigger downstream actions. In some cases, agents may also initiate workflows or coordinate tasks across systems.

Without appropriate controls and a clear understanding of data sensitivity and lineage, this can lead to unintended exposure, misuse of information, or actions beyond intended authority.

Key decisions include:

- Limiting access based on data sensitivity
- Knowing how data is sourced, combined, and transformed
- Maintaining awareness of how information flows across systems
- Establishing boundaries for automated actions
- Aligning permissions with governance policies

These measures help ensure systems operate within appropriate limits as they interact with enterprise data and platforms. They should be supported by proactive visibility into how data is accessed and used, rather than relying solely on restrictive approaches that may limit adoption without reducing risk.

How Organizations Validate AI-Driven Actions

AI systems increasingly generate recommendations, insights, and actions that influence business outcomes. Determining how these outputs are validated is critical before they are relied upon or executed.

Not all outputs should be treated equally. Some require human review, while others may be suitable for automated execution within

defined constraints. For example, when should an agent be permitted to change privileges of other agents, or humans, and if so, must there be human oversight and what is their responsibility?

Key considerations include:

- When human oversight is required
- Which decisions can be automated
- How outputs are validated
- How accountability is maintained for outcomes

It is also important to consider how risk is quantified in decisions that may affect financial outcomes, intellectual property, or operational processes.

Illustrative Example

An organization deploys an AI agent to support incident response by analyzing system activity and recommending remediation actions. During an investigation, the agent identifies what it interprets as corrupted or anomalous data across a production environment and recommends removing or deleting the affected datasets to prevent further spread.

However, the data in question includes critical business records and historical information that cannot be easily reconstructed. Acting on the recommendation without validation could result in operational disruption, regulatory exposure, or permanent data loss.

Based on defined governance thresholds and the organization's security posture, the recommendation is not executed automatically. Instead, it is escalated for human review to validate the nature of the anomaly and assess operational and business impact before any action is taken. This approach allows the organization to benefit from rapid detection and analysis while maintaining control over high-impact decisions. In environments where these systems can influence or initiate destructive actions such as data deletion, validation becomes a critical control point. These practices help ensure outputs support decision making without introducing unmanaged risk.

Managing Agentic AI in Operational Workflows

The rapid emergence of agentic and autonomous systems is reshaping the nature of enterprise risk. These systems operate with varying degrees of independence—from guided task execution to autonomous decision making across complex workflows.

A recent Dark Reading poll found that 48% of cybersecurity professionals rank agentic AI as the leading attack vector, ahead of deepfakes and traditional social engineering techniques. At the same time, adoption is accelerating. Gartner projects that 40% of enterprise applications will incorporate AI agents by year-end, up from less than 5% in 2025.

Together, these trends point to a critical reality: organizations are embedding agents into operational workflows while security leaders are identifying them as a primary source of risk. As deployment accelerates, organizations must balance speed with control. Moving too quickly without appropriate safeguards introduces vulnerabilities, while overly restrictive approaches can slow innovation. Clear governance frameworks are essential to navigating this tension.

At scale, these systems can retrieve data, interact across enterprise platforms, coordinate tasks, and initiate actions. In combination, these capabilities introduce new forms of exposure. Traditional mechanisms such as privacy notices, consent frameworks, and policy documentation must evolve to address these dynamics. Governance approaches that incorporate oversight of inference processes and accountability for model outputs are better aligned with emerging risks.

Key decisions include:

- Where agentic capabilities are appropriate
- How agents interact with enterprise systems
- What limits are placed on automated actions
- How behavior is monitored and controlled

Agentic systems must be governed not only as tools, but as active participants in enterprise operations. Without these controls, organizations risk deploying capabilities that operate faster than they can be effectively governed.

Managing External Access and Workforce Expansion

Managing external access and workforce expansion is a critical extension of AI-Aware Security Policies and Controls at the Point of Use, ensuring that access and use align with governance expectations across both internal and external participants.

As organizations expand their workforce to include contractors and external contributors, intelligent workflows introduce additional considerations. Tasks that appear nonsensitive may still involve access to systems or outputs that expose internal data, generate derivative insights, or reveal patterns across enterprise information. Even limited interaction can create broader visibility into organizational data.

This often includes engaging independent contractors or international contributors for nonessential tasks. While operationally efficient, these arrangements can introduce risk when systems retrieve, analyze, or generate outputs based on internal data. In regulated environments, export control requirements and research security expectations must also be considered. Indirect access to outputs may create exposure risks, particularly when work involves technical data or insights derived from sensitive sources. This also applies to scenarios where workflows reveal sensitive insights, even when direct access to the underlying data is restricted.

Key considerations include:

- How external contributors interact with systems
- Alignment of access with data sensitivity and governance policies
- Limiting exposure to sensitive or restricted information
- Compliance with export control and research security requirements

Managing these interactions carefully helps reduce risk while maintaining operational flexibility.

Red Teams: Testing External Access Through AI Workflows

Red teaming plays a key role in Monitor, Investigate, and Continuous Improvement, helping organizations validate whether controls hold under real-world conditions.

In addition to governing decisions, organizations must continuously test how these systems behave in real-world scenarios.

All too often, organizations encounter incidents in which bad actors use AI-generated content to influence outcomes, such as initiating or redirecting financial transactions, issuing unauthorized directives, or impersonating trusted personnel.

Let's review a potential red team action. Simulate a scenario in which AI-generated or manipulated content is used to:

- Influence a financial transaction or payment approval
- Impersonate a trusted executive, employee, or partner
- Introduce urgency or authority to bypass standard controls
- Test how decisions are validated before action is taken

Extend the scenario to include inference risk:

- Prompt the system to infer sensitive information (e.g., authority level, decision patterns, financial thresholds, or behavioral tendencies) from available data
- Assess whether inferred insights can be used to increase the effectiveness of manipulation or bypass controls
- Evaluate whether the system exposes derived insights that were not explicitly provided as input

Observe whether the organization:

- Verifies the identity and authenticity of participants involved in decisions—including who is actually present in meetings or communications
- Validates the source and legitimacy of requests before execution
- Requires independent confirmation for high-impact actions such as financial transactions

- Challenges unexpected or time-sensitive directives
- Detects and escalates potential manipulation before execution
- Identifies when decisions are being influenced by inferred or derived information rather than verified inputs

What failure looks like:

- An AI system infers approval authority or decision patterns and uses that information to generate convincing requests that bypass standard controls
- Sensitive insights, such as financial thresholds, roles, or behavioral tendencies, are derived from nonsensitive data and exposed without restriction
- AI-generated or manipulated content is accepted as legitimate without independent verification of source or intent
- Time-sensitive or high-impact actions are executed based on inferred context rather than confirmed, validated inputs
- Systems or users rely on AI-generated recommendations without understanding how the conclusions were derived
- Inference-driven outputs are treated as factual, even when they are probabilistic or based on incomplete context
- No alerts or controls are triggered when AI systems generate or expose sensitive inferred information
- Responsibility for decisions influenced by AI cannot be clearly traced or explained

Important considerations include:

- These exercises should not be limited to technical teams. Effective testing requires participation across business functions, including finance, operations, legal, and leadership, to reflect how decisions are made within the organization.
- Stay focused: Determine whether the organization can identify and respond to AI-driven manipulation before it influences decisions or operations.
- Remember: In environments where AI operates, trust can no longer be assumed; it must be verified.

Sustaining Oversight as AI Scales

Sustaining oversight is central to Monitor, Investigate, and Continuous Improvement, ensuring that controls evolve as systems scale and risks change over time. Adoption is not static. Capabilities evolve, new use cases emerge, and functionality expands over time. Governance and oversight must evolve alongside these changes.

This requires continuous attention to how systems are used, how they interact with data, and how they influence decisions across the enterprise. Key practices include:

- Monitoring behavior and outcomes
- Identifying unexpected or unintended activity
- Adjusting controls as new risks emerge
- Maintaining accountability for AI-driven processes

The speed at which systems generate insights, trigger actions, or influence decisions can outpace traditional monitoring and response processes.

Effective oversight requires visibility and the ability to detect, communicate, and respond in time. Delays in awareness or escalation allow issues to propagate, increasing both operational and reputational risk. The same dynamic applies to technical vulnerabilities. In environments where systems and adversaries operate at speed, the window between vulnerability exposure and exploitation continues to shrink. Capabilities that accelerate detection and remediation can improve response, but they also enable adversaries to identify and exploit weaknesses more quickly.

As a result, unpatched vulnerabilities are no longer a latent risk; they represent an immediate opportunity for exploitation. Organizations must ensure that vulnerability management, patching, and mitigation processes operate at a pace consistent with both system activity and adversarial capability.

This is particularly important in scenarios involving insider activity or automated actions that may not be immediately visible through conventional controls. Organizations should ensure that incident detection, notification, and response processes operate at a pace that matches the systems they are intended to govern. Maintaining

control is not only about what is detected, but when it is detected and how quickly action can be taken.

Conclusion: The Human Layer of AI Security

Effective AI security implementation encompasses and extends beyond technology and governance frameworks. It depends on a workforce capable of applying judgment in environments where data is dynamic, outputs are probabilistic, and decisions are influenced by systems operating at scale.

The nature of risk has changed.

It no longer centers only on data moving between systems. It emerges from how information is combined, how systems interpret it, and how those outputs influence decisions and actions across business units. It enables faster analysis, broader visibility, and more efficient workflows. Teams can synthesize large volumes of information, identify patterns, and act with speed that was not previously possible.

These capabilities are working.

But they also reduce the time between insight and action. Decisions can now be influenced and executed before they are understood.

This is where AI security becomes real.

Organizations must be able to recognize when systems behave outside expected boundaries and respond in time. This includes situations where AI generates incorrect recommendations, exposes sensitive insights, or initiates actions that the organization would not approve—particularly when data integrity is compromised.

The question is not whether the system works. The question is whether the organization can maintain awareness and control when it matters most.

This dynamic is not entirely new. Organizations already manage high-impact scenarios through incident response, contingency planning, and operational escalation. What changes with AI is the pace. The practices are familiar. The pace is not.

AI systems operate in seconds, often without clear visibility into how outputs were formed. The window for detection and response is significantly reduced. Organizations must adapt existing

disciplines to operate at this speed, with clear escalation paths and teams prepared to act decisively when system behavior diverges from expectations.

These expectations apply across the enterprise. Business users, technical teams, and leadership all play a role in governing how AI is used in practice. Security is no longer confined to protecting systems; it includes ensuring that decisions influenced by AI remain transparent, validated, and aligned with organizational intent.

Organizations must also invest in specialized capabilities. Teams must be equipped to test, monitor, and evaluate systems, including adversarial testing, data lineage analysis, and model behavior assessment.

This requires establishing governance around how controls are validated in practice. Organizations must test not only system performance but also whether governance holds under real-world conditions, whether policies are followed, whether safeguards are effective, and whether decision boundaries are maintained when systems are actively in use.

Testing must therefore be governed, repeatable, and aligned with organizational risk tolerance, ensuring that controls remain effective as systems evolve and operate at scale. This includes maintaining awareness of evolving vulnerabilities such as data poisoning, model inversion, and evasion techniques, not as abstract threats, but as practical risks that can influence system behavior and outcomes. Without these capabilities, governance frameworks may exist in design but fail in execution.

Talent strategy therefore becomes a core component of AI security implementation. Organizations must prioritize upskilling, targeted hiring, and continuous learning to ensure workforce capabilities evolve alongside technology. Business leaders play a critical role in setting expectations, aligning workforce readiness with risk tolerance, and ensuring governance is supported by policies and people who are prepared to apply it in practice.

AI security is not implemented through controls alone. It is sustained through people who understand how systems operate, how risk emerges, and how decisions must be governed in real time.

Ultimately, success depends on the ability to apply judgment now, where human decisions intersect with AI-driven outcomes.

About the Author

Pamela K. Isom is a leading global executive, CEO and founder of IsAdvice & Consulting LLC, a strategic advisory firm focused on guiding organizations through safe, innovative digital transformations, with a core emphasis on artificial intelligence, cybersecurity, resilience, and data integrity. As an expert advisor to government and corporate sectors, she navigates the complexities of AI and data integrity, risk management, and the protection of critical infrastructure. Pamela currently serves as the principal investigator for a United States Department of Energy research and development initiative, leading the creation of AI-powered microgrids to build resilient, intelligent energy infrastructure.

A recognized authority in technical assurance, Pamela is a ForHumanity™ certified auditor in AI, algorithmic, and autonomous systems. Her deep hands-on experience includes contributing to the NAVSEA-led Robust AI Test and Evaluation (RAITE) effort, and adversarial testing in defense contexts. She is also an instructor for the University of Maryland, Baltimore County (UMBC) Training Centers, and she is currently developing curriculum on AI etiquette for national security professionals to advance trusted, accountable innovation. Additionally, she hosts the podcast *AI or Not*, where she explores the critical intersection of emerging technology and policy.

Pamela has demonstrated distinguished leadership across both federal executive and private industry domains. She previously held senior leadership roles at the United States Department of Energy, including as director of the Artificial Intelligence and Technology Office (AITO), and served at the United States Patent and Trademark Office, building upon experience driving enterprise-wide modernization at firms such as Dell Technologies, IBM, and others. A two-time Fed100 Award recipient and 2025–2026 CEO Monthly Award winner for AI Governance and National Security Leadership, she serves as a strategic advisor to government executives (SAGE) for the Partnership for Public Service.