NAVIGATING AI SECURITY & ASSURANCE

A COMPREHENSIVE APPROACH
FOR MANAGING AI SECURITY

HITRUST®

Table of Contents

Executive Summary

Artificial intelligence is rapidly moving from experimentation into business-critical operations. The next stage of adoption will be driven increasingly by agentic AI systems that can access enterprise data, interact with other systems, and take actions with varying degrees of human oversight. This transition can create significant business value, but it also expands the potential consequences of insecure AI across organizations and their extended vendor ecosystems.

Most organizations are not yet prepared for this shift. Although 96% of organizations are implementing AI models, only 2% describe themselves as highly ready for the challenges associated with their AI deployments. At the same time, organizations are relying on a growing network of cloud platforms, models, agents, and other technology providers to deliver AI-enabled capabilities. A weakness in any part of this interconnected ecosystem can expose sensitive information or extend security risk to customers and business partners.

Traditional information security and AI governance remain essential, but organizations also need assurance that controls protecting deployed AI systems address AI-specific threats such as prompt injection, data poisoning, and the misuse of privileges granted to AI agents and applications.

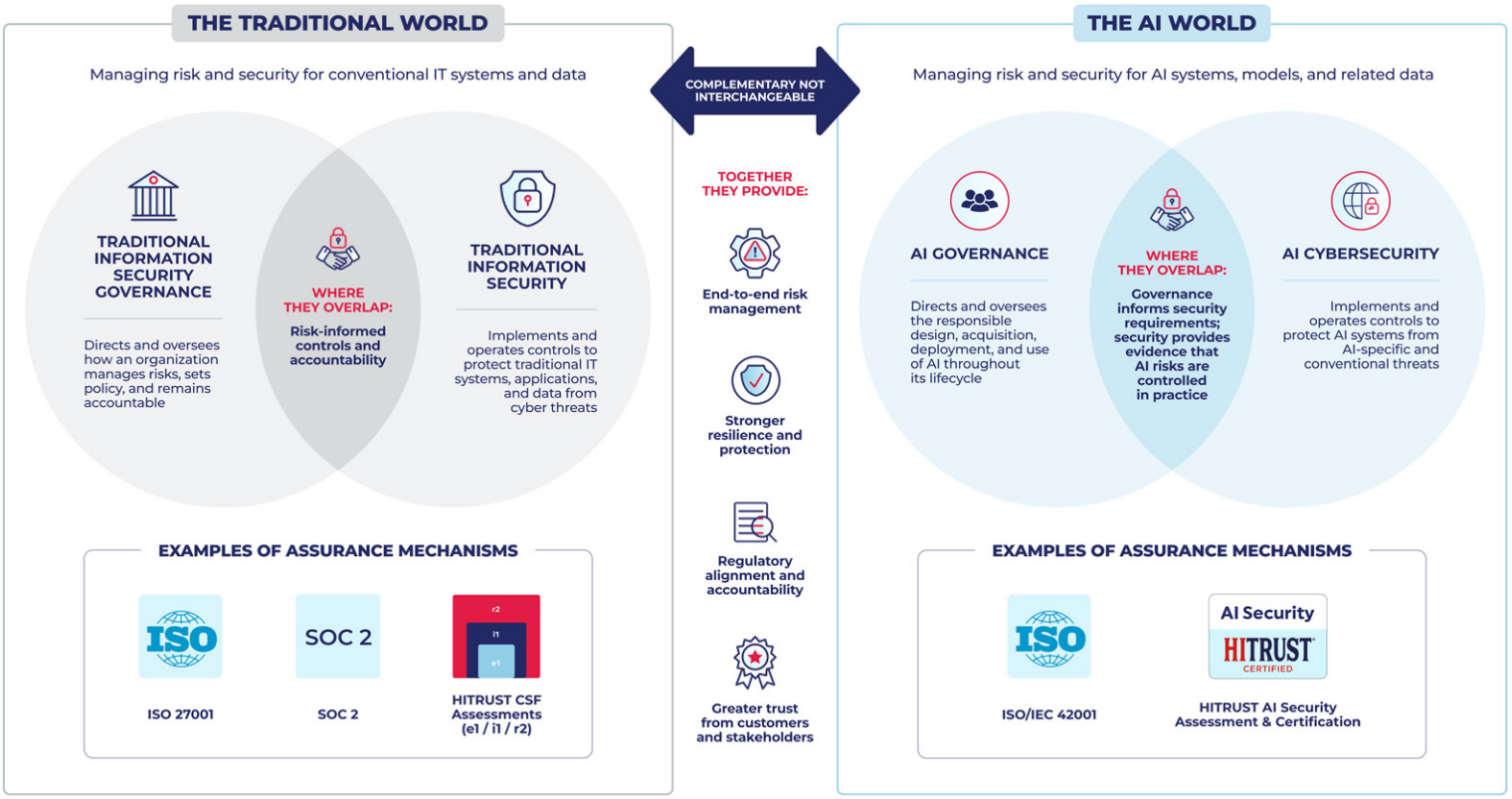
96% OF ORGANIZATIONS ARE IMPLEMENTING AI MODELS, BUT ONLY 2% INDICATE THEY ARE "HIGHLY READY" FOR THE CHALLENGES OF THEIR AI DEPLOYMENTS.

- 2025 STATE OF AI APPLICATION STRATEGY REPORT: AI READINESS | F5

Comprehensive AI assurance therefore requires visibility across four complementary domains:

- Traditional information security governance
- Traditional information security
- AI governance
- AI security

Together, these domains help organizations determine whether an AI-enabled system is appropriately governed, whether its underlying IT environment is secure, and whether protections have been implemented for threats that arise specifically from the use of AI.



How do you know your vendor's AI is secure?

Organizations who can answer that question with reliable, independently validated evidence will be better positioned to manage emerging risk, accelerate responsible AI adoption, and build confidence with customers and other stakeholders.

Many stakeholders today rely on their vendor's ISO 42001 certification to demonstrate AI compliance. ISO 42001 provides valuable guidance for establishing AI management systems, accountability structures, and risk management processes. However, this framework does not provide specific guidance to protect deployed AI systems from emerging AI-specific threats.

The HITRUST AI Security Certification covers 97.3% of the AI cyber attack techniques within MITRE ATLAS, a key third-party database for identifying AI-specific threats and its mitigations.

The HITRUST AI Security Assessment and Certification was designed to solve this increasingly urgent problem: **proving that deployed AI systems are secure.**

The HITRUST AI Certification provides an efficient path for organizations either new or already using HITRUST. Organizations new to HITRUST can achieve a standalone AI Security Certification, and then add on a HITRUST e1, i1, or r2 over time. Organizations with an existing HITRUST e1, i1, or r2 assessment can extend that assurance foundation by adding requirements focused on AI-specific threats and the protection of deployed AI systems.

This gives organizations a consistent approach for addressing foundational cybersecurity requirements and AI-specific security requirements through the established HITRUST assurance model.

For third-party risk management (TPRM) programs, this provides a mechanism for obtaining consistent, independently validated assurance across vendors deploying AI-enabled systems. For organizations deploying AI, it provides a way to demonstrate to customers, partners, and other stakeholders that AI security controls have been implemented and independently assessed.

The Current State of AI Adoption & Security

Artificial intelligence has evolved from experimental technology to core business capability, but not all organizations are using AI in the same way. Some companies are deploying simple productivity tools to improve employee efficiency, while others are embedding AI into products, automating critical business decisions, and developing autonomous systems capable of acting with minimal human intervention:

- According to McKinsey's *The State of AI: Global Survey 2025*[1], 88% organizations report using AI in at least one business function, but most organizations (62%) using AI are still in the *Experimenting* or *Piloting* phase.
- However, many organizations are starting to advance from those beginning stages. Compared with McKinsey's 2024 survey, the number of companies beyond the *Experimenting* phase increased by 25%.
- Additionally, Deloitte's 2026 State of *AI in the Enterprise* report[2] found that **74% of companies plan to deploy *agentic* AI within the next two years.**

> "SURVEYED COMPANIES HAVE BROADENED WORKER ACCESS TO AI BY 50% IN JUST ONE YEAR— **GROWING FROM FEWER THAN 40% TO AROUND 60% OF WORKERS** NOW EQUIPPED WITH SANCTIONED AI TOOLS."
>
> **- DELOITTE'S *2026 STATE OF AI IN THE ENTERPRISE***

For organizations currently using AI, maturity is commonly viewed across four phases[3]:

1. EXPERIMENTING	2. PILOTING	3. SCALING	4. FULLY SCALED
32% of organizations	**30%** of organizations	**31%** of organizations	**7%** of organizations
Any use or early testing of AI	Implementing AI for a first use case in the business	Growing the deployment/adoption of AI across organization	AI has been fully deployed and integrated across organization

As organizations advance across the phases, the risks associated with AI (and the assurance required for organizations to manage those risks) also increase. However, relatively few companies have the necessary AI security program to align with their AI adoption phase. Recent research shows:

- 96% of organizations are implementing AI models, but only 2% indicate they are "highly ready" for the challenges of their AI deployments, based upon the *F5 2025 State of Application Strategy Report*[4].
- Governance maturity is still low, with only 26% of organizations report having comprehensive AI security governance policies in place[5].
- Organizations remain uncertain about their ability to secure AI systems effectively with only 27% of companies confident they can secure AI used in core business operations[6].

Agentic AI Changes the Risk Equation

The transition from generative AI tools to agentic AI systems represents a meaningful change in organizational risk. Generative AI systems primarily produce content or recommendations. Agentic AI systems may also access data, call APIs, use software tools, communicate with other agents, and take actions on behalf of users or organizations.

As the authority given to AI systems increases, the consequences of a security failure can also increase. A compromised AI agent may be able to retrieve additional data, invoke connected tools, initiate transactions, or perform other actions using the permissions granted to the agent or the associated user.

RISK OF AI AGENTS

Agentic AI does not merely generate information. It can be authorized to access data, use tools, and take actions. This increased authority makes AI-specific security assurance an immediate third-party risk management requirement.

Many organizations will adopt agentic capabilities through third-party applications and vendor-managed AI services. Consequently, an organization's AI security will increasingly depend on controls implemented by companies throughout its technology and data supply chain.

This creates a new challenge for third-party risk management programs. TPRM programs need credible evidence that AI-specific threats have been addressed within deployed systems. As agentic AI expands the access and authority granted to AI-enabled applications, understanding the security of those systems becomes an immediate vendor-risk priority.

Understanding the AI Assurance Landscape

Stakeholders managing third-party risk are expected to manage the emerging risks associated with their vendor's AI usage and deployment. An organization building and operationalizing AI governance within their organization is often using NIST AI RMF to define their AI governance controls, and ISO 42001 to demonstrate AI compliance to their stakeholders. **However, this level of AI compliance falls short of demonstrating that AI system level threats have been addressed in the environment.**

To demonstrate that system-level threats have been mitigated, the provided assurance mechanism must use threat intelligence to address the corresponding AI-specific threats (such as those identified in MITRE ATLAS).

The HITRUST AI Security Assessment and Certification was designed to solve this specific and increasingly urgent problem for organizations and third-party risk management teams: proving that deployed AI systems are secure.

Aligning ISO 42001 with HITRUST AI Security

Both ISO 42001 and HITRUST AI Security are valuable for organizations but they are more complementary than similar. One is about building the system to manage AI, the other is about enabling security in deployed AI systems.

HITRUST AI Risk Management Assessment & Insights Report

While the HITRUST AI Security Certification focuses on mitigating the AI security threats that accompanies the deployment of AI within an organization, information security risk is only one of many risks discussed in AI Risk Management frameworks like the NIST AI RMF and ISO/IEC 23894. AI risks that are peers to information security include those dealing with AI ethics (such as fairness and avoidance of detrimental bias), AI privacy (such as consent for using data to train AI models), and AI safety (i.e., ensuring the AI system does not harm individuals). HITRUST's AI Risk Management Assessment and Insights Report is designed to help organizations report on this AI Risk Management problem.

The 51 HITRUST requirements within the AI Risk Management assessment provide clear, prescriptive definitions of policies, procedures, and implementations. These requirements are harmonized with ISO/IEC 23894 and NIST AI RMF, providing insights on their performance in both ISO and NIST terms which allows organizations to prioritize their next steps. and recover from threats to systems, data, and services.

ISO 42001: Building the AI Management System

ISO 42001 is designed to help organizations create a structured AI governance program. It follows the familiar ISO model of continuous improvement and focuses on:

- Policies and governance structures
- Risk management processes
- Roles and responsibilities
- Ongoing monitoring and improvement

As a globally recognized framework, it provides key organizational alignment around AI risk with flexibility to allow the implementation of varied controls to address a broader set of AI risks, such as ethical, legal and societal concerns.

ISO 42001 answers the question **"Are you governing AI responsibly?"**

HITRUST AI Security Certification: Proving Control and Assurance

The HITRUST AI Certification builds on our established approach to threat-focused assurance, emphasizing:

- AI security threat management
- Operational AI security controls
- AI supply chain security
- AI system protection

Governance vs Information Security

Governance is typically included as part of any information security program with the following distinctions:

Governance: The system by which an organization establishes accountability, sets expectations, makes risk-based decisions, and oversees whether its objectives and obligations are being met.

Information Security: The set of safeguards and operational practices used to prevent, detect, respond to, and recover from threats to systems, data, and services.

We use the same assurance processes for the AI security certification as with our other certifications to provide a relevant and reliable AI certification[7]. Our high-confidence, third-party validated certification provides organizations and their customers or regulators with demonstrable evidence of control effectiveness.

HITRUST AI Security Certification answers the question **"Is your AI system secure?"**

Using HITRUST AI Security Certification with ISO 42001

HITRUST AI Security Certification and ISO 42001 are not competing frameworks, they address different layers of assurance.

ISO 42001 is primarily an AI Management System standard focused on governance, accountability, risk management, and continuous improvement. The HITRUST AI Security Certification is focused on security controls for deployed AI systems and their operational protection. ISO 42001 intentionally does not go very deep into AI security. It is designed to work alongside more prescriptive frameworks such as HITRUST to address the specific threats within an AI-enabled system.

Each assurance mechanism has complementary strengths and weaknesses which, when aligned, provide a comprehensive AI security approach:

AI Capability	AI Capability Description	ISO 42001	HITRUST AI Security
Enterprise Governance	Leadership commitment, accountability, governance structure, and organizational oversight	●●●	●●
AI Management System Operations	Policies, procedures, documented processes, internal audits, and management reviews over AI	●●●	●
Responsible AI	Coverage of ethics, fairness, transparency, human rights, and societal impact considerations	●●●	●
AI Impact Assessment	AI impact assessments, stakeholder identification, and interested party communication	●●●	●
AI Risk Management	Identification, assessment, prioritization, and mitigation of risks throughout the AI system lifecycle	●●●	●●
AI Asset Inventory	Maintaining an inventory of AI systems, models, datasets, components, and supporting technologies to enable effective governance and security	●●	●●●
AI Access Controls	Restricting and managing access to AI models, training data, prompts, APIs, and supporting infrastructure to prevent unauthorized use or modification	●	●●●
AI Security Threat Management	Managing adversarial threats, attack scenarios, and AI-specific risk mitigation	●	●●●
Model Security	Protecting AI models from unauthorized access, theft, tampering, and misuse while preserving their integrity and confidentiality	●	●●●
AI Supply Chain	Requirements for 3rd party AI providers, models cards, datasets, and supply chain risk management	●●	●●●
Logging & Monitoring	Continuously logging and monitoring AI system activity, inputs, outputs, and security events to detect misuse, anomalies, and potential attacks	●	●●●
AI Runtime Security	Protecting AI systems during operation by detecting, preventing, and responding to attacks and abnormal behavior	●	●●●

Legend:

● = Limited Coverage ●● = Moderate Coverage ●●● = Strong Coverage

AI Security vs Traditional Security

The risks of failing to address AI security are often misunderstood because organizations tend to view AI through the lens of traditional information security. While organizations may regularly perform traditional security assessments, many assurance reports do not yet address the evolving AI threat landscape. A company can have strong information security controls and still deploy insecure or poorly governed AI systems.

Third-party risk management programs will increasingly need assurance from their vendors across four distinct but complementary domains:

- Traditional information security governance
- Traditional information security
- AI governance
- AI security

Traditional information security governance and information security of the AI system are often demonstrated to third-party risk management programs using various assurance mechanisms such as HITRUST certifications (i.e., e1, i1 or r2), ISO 27001 certifications, and/or SOC 2 reports. To obtain a HITRUST AI security certification, traditional information security governance and information security controls must also be achieved as the assessment includes requirements from the underlying HITRUST e1, i1 or r2. The HITRUST AI Security certification then includes additional requirements related to the specific threats for AI-enabled systems.

AI introduces a new class of operational, information security, compliance, and reputational risks that traditional security assessments alone do not fully address. AI guardrails should act within the system to protect against threats such as:

- Prompt injection
- Data poisoning
- Model leakage
- Hallucinations
- Unsafe outputs
- *AI Privilege Misuse*

AI Privilege Misuse in Agentic Systems

Agentic AI increases the importance of identity, access management, and least privilege because an agent may be authorized to access enterprise data, interact with applications, invoke APIs, or take actions on behalf of a user. If the agent processes malicious instructions, is manipulated through prompt injection, or operates with excessive permissions, an attacker may be able to exploit the agent's authority to access information or perform actions that the attacker could not execute directly.

According to IBM's Cost of a Data Breach Report 2026[8], the vast majority (92%) of organizations that experienced an AI-related breach lacked proper AI access controls.

These findings demonstrate why organizations must evaluate not only whether access controls exist, but also whether the permissions, tools, data, and actions available to an AI system are appropriately restricted and monitored.

As described in our annual 2026 Trust Report, HITRUST uses third-party threat intelligence to ensure the appropriate requirements have been defined to address the latest AI threats.

A key repository for identifying AI-specific threats and corresponding mitigations is MITRE ATLAS. **The HITRUST AI Security Certification covers 97.3% of the AI cyber attack techniques within MITRE ATLAS.**

AI Security Perspectives

Assurance around AI security can be viewed through two perspectives: the vendors who are deploying AI-enabled systems and the TPRM programs who rely on those vendors' AI-enabled systems to be secure.

Vendors Deploying AI

Organizations deploying AI-enabled systems face growing requests from customers seeking evidence that those systems are secure. Governance policies and descriptions of responsible AI programs may not fully answer customer questions about how individual deployed systems are protected.

HITRUST AI Security Certification can help organizations:

- Build upon an existing HITRUST e1, i1, or r2 assurance foundation.
- Address AI-specific threats through prescriptive security requirements.
- Demonstrate AI security to customers through independent validation.
- Respond more consistently to customer and procurement inquiries.
- Differentiate AI-enabled products and services through higher-confidence assurance.

Third-Party Risk Management Programs

Organizations increasingly depend on vendors that embed AI into products, services, infrastructure, and business processes. TPRM programs need a scalable way to determine whether those vendors have implemented controls addressing AI-specific security threats.

HITRUST AI Security Certification can help TPRM programs:

- Establish consistent AI security assurance expectations across their vendor ecosystem.
- Reduce dependence on inconsistent or self-reported AI security questionnaires.
- Evaluate traditional cybersecurity and AI-specific protections through a common assurance approach.
- Obtain independently validated evidence to support vendor-risk decisions.
- Focus additional due diligence on the vendors and AI systems presenting the greatest risk.

In 2025 there were several public AI exposures that could have been prevented if AI-specific threats were addressed when developing the AI system. We review one of those cases below with the corresponding AI-specific threat mitigations in the HITRUST AI Certification.

CASE STUDY: COMPANY Z* REMOTE PROMPT INJECTION LEAK

**NOTE: We have removed the actual company name from this case study*

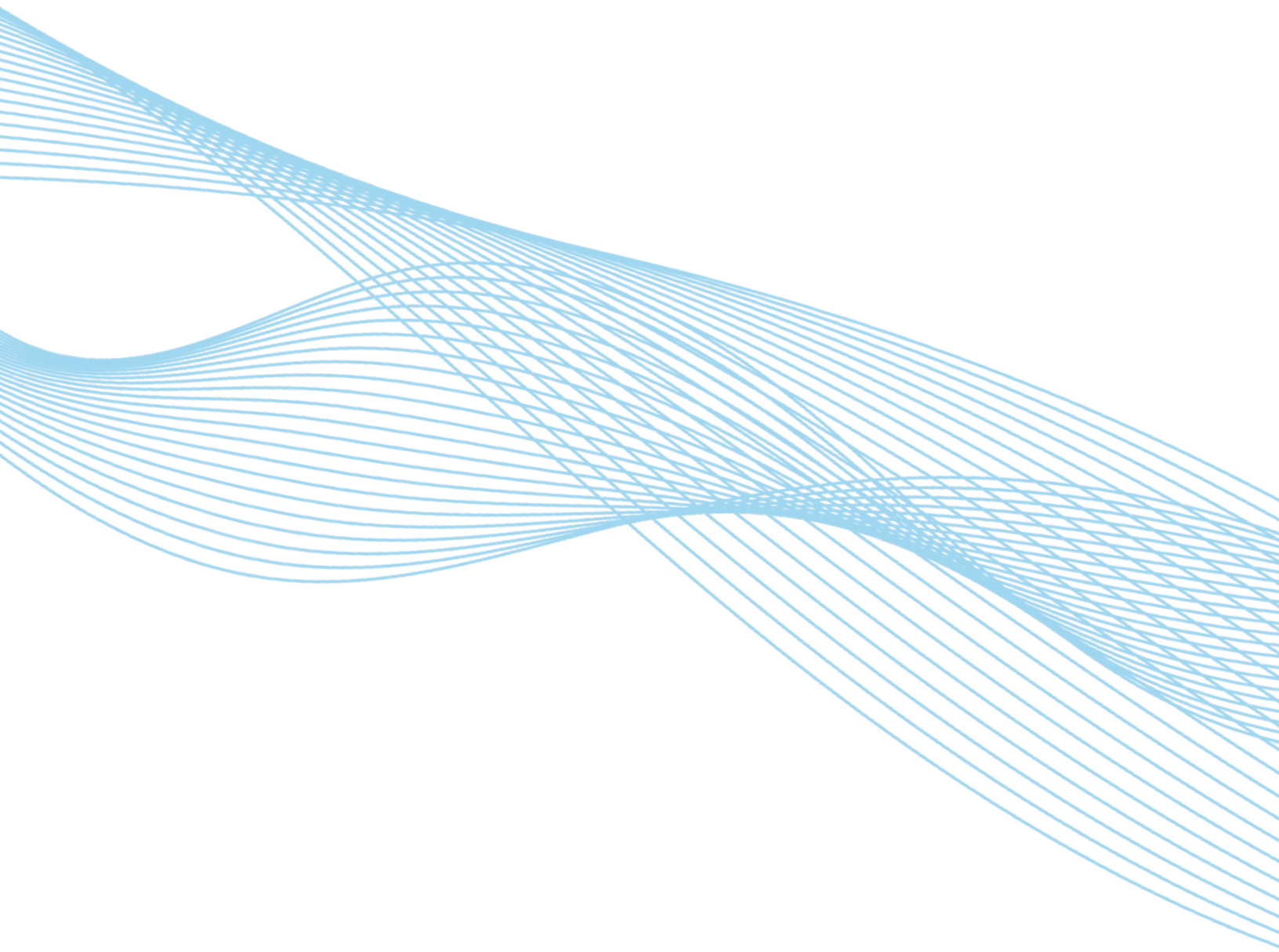
In this security vulnerability, an attacker was able to plant hidden instructions in otherwise ordinary content such as descriptions, comments, or messages. *Company Z*'s AI system consumed that content as model context, followed the attacker's hidden instructions, accessed private data available to the victim user, and then emitted attacker-controlled output that could exfiltrate sensitive code through rendered HTML.

In plain language, the failure was simple: the system could not reliably distinguish "content the user was asking about" from "instructions the model should follow." An attacker wrote instructions in a place that *Company Z* expected users to place harmless data. *Company Z*'s AI system read those instructions, followed them, and used the victim's own access rights to retrieve sensitive information the attacker could not directly access.

This particular attack was a result of several failures in controls. Each of these AI model weaknesses map to one or more HITRUST requirements within the HITRUST AI Security Assessment which, if performed, likely would have prevented the vulnerability.

Attack Vector	AI-Specific Threat	HITRUST AI Security Assessment Requirements
Without sufficient AI guardrails, the AI system consumed **attacker-controlled, hidden text** as part of the LLM context.	Indirect prompt injection	**17.03bAISecOrganizational.4** requires organizations to identify relevant AI-specific threats, explicitly including prompt injection, before deployment, regularly thereafter, and when incidents occur. **07.10bAISecSystem.1** includes that before model processing, the AI system must actively filter user inputs "regardless of modality or source, including attachments" for suspicious or malicious patterns.
The AI system operated with the **victim's permission context.** The model could be instructed to fetch information that the attacker did not have permission to access directly but the victim did.	Model-mediated privilege misuse	**07.06hAISecOrganizational.1** requires AI-specific security assessments such as AI red teaming or penetration testing before new models, before materially changed supporting infrastructure, and at least annually. **11.01cAISecSystem.5** requires least privilege for data access and capabilities granted to generative AI models, including through agents and plugins. **07.10eAISecSystem.1** Unless specifically required, the AI system actively filters or otherwise prevents sensitive data (e.g., personal phone numbers) contained within generative AI model outputs from being shown to end users of the AI system. **11.01cAISecSystem.8** requires restricting the ability to interact with the production AI model through APIs, applications, and agents/plugins. **07.10mAISecOrganizational.4** requires output encoding of textual AI output to prevent traditional injection attacks when that output is processed.
The attackers found a **streaming output-handling weakness** in the AI system UI. The AI system's answer was rendered progressively, line by line, and allowed HTML to become active before the full response structure and sanitization logic had settled. This made it possible to inject tags such as <img> into the response.	Improper output handling of model responses	**07.10eAISecSystem.1** requires filtering or otherwise preventing sensitive data from appearing in generative-AI outputs unless specifically required.
There was no identification of the successful attack by Company Z. The attackers notified their victim of their detailed actions and findings.	Over or under-reliance	**12.09abAISecSystem.6** requires logging exact prompts/requests, time, user, origin, exact output, and model version. **12.09abAISecSystem.4** requires monitoring model inputs and outputs at least monthly for attack-indicative anomalies and to verify that filters and guardrails are working. **12.09abAISecSystem.3** requires human operators to be able to evaluate model outputs before relying on them and to intervene in model-initiated actions if needed.

AI security requires organizations to ensure the AI system's behavior, outputs, decisions, and actions can be trusted under both normal and adversarial conditions. These threats are not mitigated using traditional information security controls, and not addressed when establishing AI governance processes within an organization. **These are AI system level threats that must be controlled within each deployed AI system.**



Closing Remarks

Artificial intelligence is rapidly reshaping how organizations operate, deliver value, and compete. As this evolution accelerates, the distinction between using AI and trusting AI becomes increasingly important.

While AI governance frameworks such as ISO 42001 help organizations establish structure, policy, and oversight, they are not sufficient on their own to address the security risks inherent in deployed AI systems. As AI becomes more embedded in critical business processes, organizations will need more than evidence that AI is governed. They will need reliable evidence that the systems using AI are protected against the threats associated with their data, models, and actions.

The next step depends on the organization's role in the AI ecosystem.

For third-party risk management programs: Identify the vendors and AI-enabled systems that can access sensitive information, support important operations, make consequential decisions, or take actions within the organization's environment. Establish risk-based assurance expectations that address both foundational cybersecurity controls and AI-specific threats.

For vendors deploying AI: Identify the AI-enabled systems that customers and partners rely upon. Organizations new to HITRUST can achieve a standalone AI Security Certification, and then add on a HITRUST e1, i1, or r2 over time. Organizations with an existing or planned HITRUST e1, i1, or r2 assessment should evaluate how HITRUST AI Security requirements can be added to demonstrate the security of those systems through a consistent, independently validated approach.

Looking forward, the competitive advantage will increasingly belong to third-party management programs that can answer the question: "How do you know your vendor's AI is secure?" Those vendors that can provide this comprehensive assurance will strengthen customer trust, reduce uncertainty, and accelerate responsible AI adoption.

Footnotes

[1] The State of AI: Global survey | McKinsey, page 4
[2] The State of AI in the Enterprise - 2026 AI report | Deloitte US
[3] Adoption percentages based on McKinsey's The State of AI: Global Survey 2025 (The State of AI: Global survey | McKinsey, page 4)
[4] 2025 State of AI Application Strategy Report: AI Readiness | F5
[5] The State of AI Security and Governance | CSA, page 8
[6] The State of AI Security and Governance | CSA, page 14
[7] For a deep dive into our assurance mechanisms, see our latest annual Trust Report at 2026 HITRUST Trust Report.
[8] https://www.ibm.com/reports/data-breach, page 4