AI, Automation and the Next Generation of Technology Risk: From Prompt Injection to Agentic Security

Turning Emerging AI Threats into Resilience Opportunities

THOUGHT LEADERSHIP PAPER

MAY 2026

Table of Contents

Executive Framing

Artificial Intelligence is rapidly reshaping how organizations operate, compete, and deliver value. Across industries, we are seeing AI embedded into core business processes, from customer interaction and decision support to software development and operational automation. However, this transformation is not simply expanding the digital attack surface; it is introducing an entirely new execution layer, one where systems can reason, generate outputs, and increasingly act autonomously.

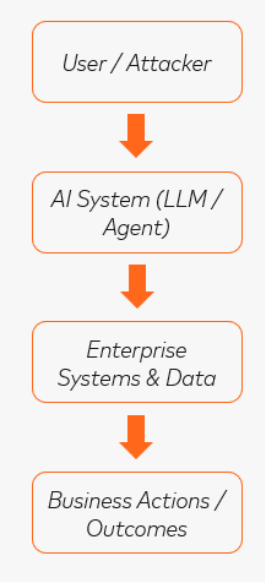
What makes this shift particularly significant is the growing gap between adoption and control. Organizations are scaling AI capabilities at speed, often ahead of their ability to fully understand, govern, and secure them. Industry research indicates that enterprise AI usage has surged significantly over the past year, while visibility into how these systems are accessed, integrated, and exposed remains limited. At the same time, security leaders increasingly acknowledge that they are not fully prepared to manage AI-specific risks, highlighting a structural misalignment between innovation and governance.

AI Adoption vs AI Security Readiness

"Organizations are scaling AI faster than they can securely govern it."

This creates a fundamental shift in the nature of technology risk. Traditional cybersecurity models were designed to protect deterministic systems; systems that behave predictably, execute predefined logic, and can be tested against known conditions. AI systems, by contrast, operate through probabilistic reasoning and dynamic context interpretation. Their behavior is shaped not only by code, but by data, prompts, external integrations, and user interaction patterns. As a result, they introduce risks that are less about software vulnerabilities and more about the manipulation of system behavior itself. [2, 9]

AI as an Execution Layer



As organizations transition into this new paradigm, it becomes clear that this is not an incremental evolution of existing risk—it is a structural transformation. The nature of exposure has shifted from exploiting code and configurations to influencing how systems interpret, reason, and act. This introduces new forms of non-linear and context-dependent risk that traditional control frameworks were not designed to address. [2, 11]

Evolution of Technology Risk

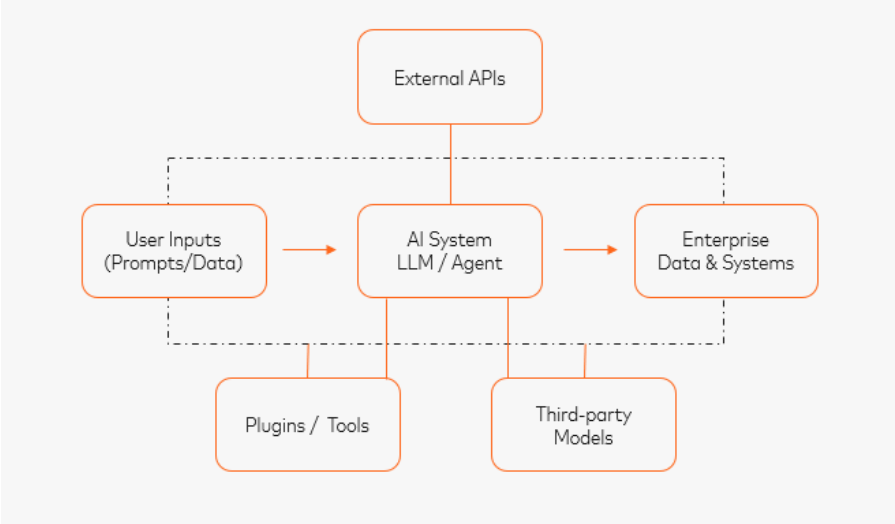
Era	System Type	Security Focus	Primary Risk
Traditional IT	Deterministic	Perimeter Security	Exploits
Cloud / Digital	Distributed	Configuration Control	Misconfigurations
AI Era (Today)	**Probabilistic**	**Behavior Governance**	**Manipulation**

"AI introduces behavior-driven risk, not just code-level vulnerabilities."

These risks are no longer theoretical. The emergence of AI-specific attack techniques—formalized through frameworks such as the OWASP Top 10 for Large Language Model Applications; demonstrates how attackers can exploit model behavior rather than underlying code. Real-world incidents reinforce this shift. The "Reprompt" attack showed how a single crafted interaction could bypass AI guardrails and extract sensitive information through indirect prompt manipulation. In parallel, recent threat disruption efforts have revealed how adversaries are leveraging AI platforms to conduct coordinated influence operations and large-scale social engineering campaigns, using AI not only as a target but as an enabler of malicious activity.

The expansion of AI capabilities is further amplified by the growing use of connectors, plugins, and external integrations, allowing AI systems to interact with enterprise data, third-party services, and operational tools. Emerging architectural patterns—including model-to-system connectivity frameworks such as MCP (Model Context Protocol) are accelerating this integration. While these developments unlock significant business value, they also introduce new trust boundaries and dependencies that are often insufficiently governed, increasing the potential impact of compromise. [3, 14]

AI Ecosystem / Integration Risk



"Each integration introduces a new trust boundary and potential attack path."

From a regulatory perspective, expectations are also evolving. Across jurisdictions, regulators are placing increasing emphasis on operational resilience, accountability, and the ability to demonstrate control over technology risk, including AI-driven processes. Organizations are therefore not only expected to deploy AI securely, but to evidence how these systems are monitored, validated, and governed within broader risk and compliance frameworks.

In this environment, cybersecurity itself is evolving. We are observing a shift from periodic control validation toward continuous risk sensing, recognizing that static assessments cannot adequately address dynamic, adaptive systems. Similarly, the focus is moving from breach prevention alone to resilience engineering—designing systems and operating models that can withstand, adapt to, and recover from disruption.

This evolution is driving the emergence of a new paradigm: agentic security. In this model, security capabilities are designed not only to protect systems, but to continuously monitor, validate, and respond to the behavior of intelligent technologies operating at machine speed.

For executive leadership, this introduces a fundamental shift in perspective. It is no longer sufficient to ask whether systems are secure in a traditional sense. The more critical question is whether organizations can effectively govern technologies that are capable of reasoning, generating outputs, and acting on behalf of users.

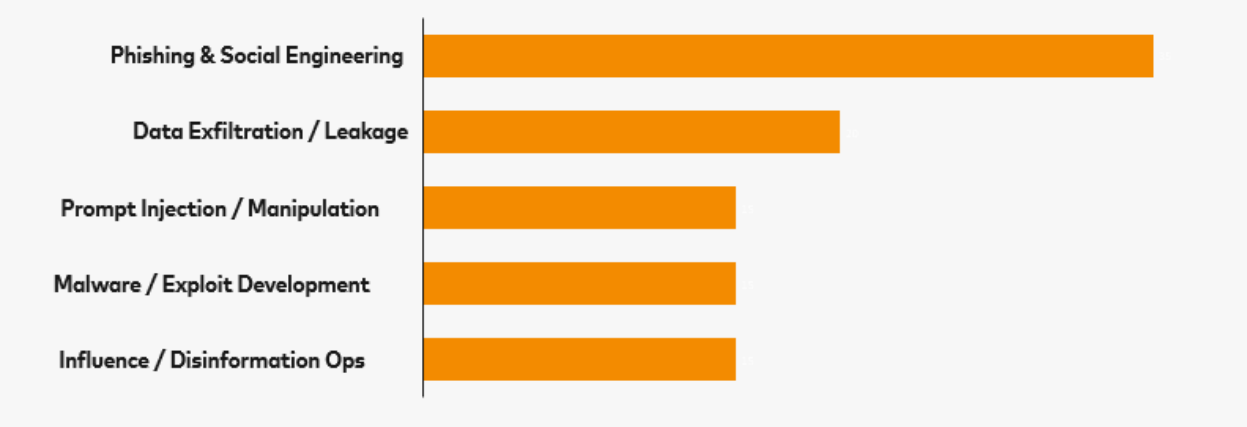
The question is no longer whether organizations will deploy AI; but whether they can effectively govern systems that increasingly reason, generate outputs, and act on behalf of users. [2, 11]

The 2026 AI Threat Landscape

Enterprise AI adoption has reached unprecedented levels. Across industries, we are observing rapid integration of generative AI into core workflows, including software development, customer engagement, analytics, and operational automation. At the same time, a significant portion of AI usage is occurring outside formal governance structures, as employees increasingly rely on external tools and platforms without centralized visibility or control. This rise of "shadow AI" is expanding the attack surface in ways that are not yet fully understood, creating blind spots in both security monitoring and risk management.

At the same time, the threat landscape is evolving at an equally accelerated pace. Threat actors are actively leveraging AI to enhance traditional attack techniques, including phishing, social engineering, reconnaissance, and exploit development. Capabilities that once required advanced technical expertise are becoming increasingly accessible, allowing attackers to scale operations with minimal cost and effort. This democratization of attack tooling is fundamentally changing how cyber threats are executed.

AI-Enabled Threat Landscape (2026)



"Threat actors are leveraging AI across multiple attack vectors, with social engineering and data exploitation leading the landscape."

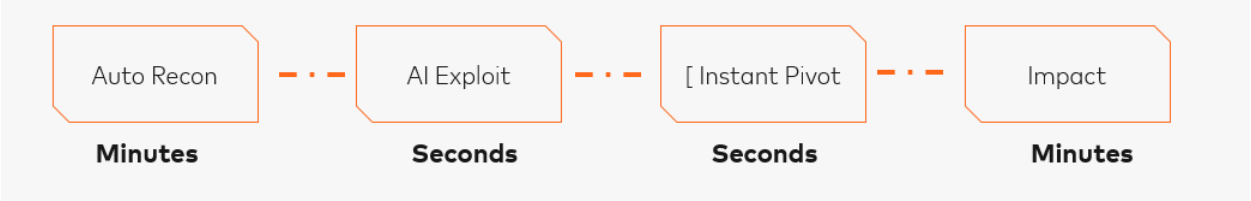
A defining characteristic of this shift is the compression of attack timelines. AI enables adversaries to automate multiple stages of the attack lifecycle, reducing the time required to identify targets, generate attack content, and execute malicious actions. As a result, attacks that previously unfolded over days or weeks can now occur within minutes or seconds, significantly reducing the window available for detection and response. [9, 15]

Attack Speed Compression

Traditional Attack Lifecycle

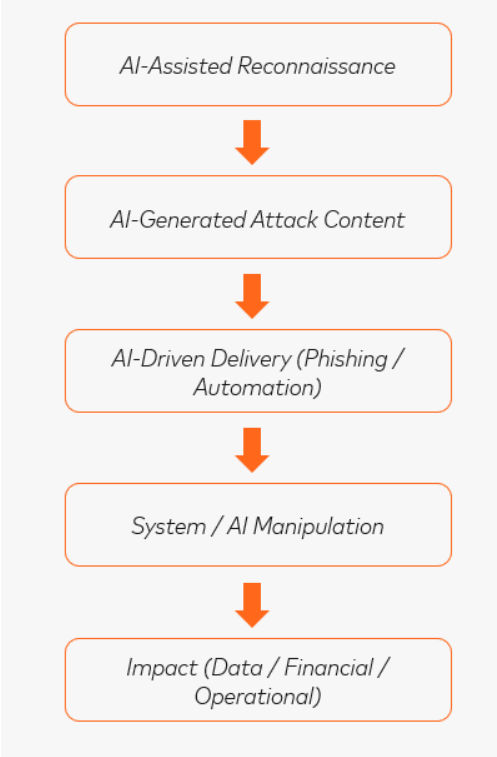
"AI compresses attack timelines from days/weeks to minutes/seconds." [12, 15]

AI-Enabled Attack Lifecycle



This acceleration is not only increasing the frequency of attacks, but also fundamentally altering their structure. Rather than executing isolated steps, attackers are increasingly chaining AI capabilities together—combining automated reconnaissance, content generation, and delivery mechanisms into streamlined, end-to-end attack workflows. [2, 8]

AI Attack Chain



"AI enables attackers to compress and integrate the full attack lifecycle into a continuous, automated workflow."

Real-world incidents are already demonstrating how these dynamics materialize in practice. The "Reprompt" attack showed how a single crafted interaction could manipulate an AI system into bypassing its own safeguards and disclosing sensitive information. Rather than exploiting code-level vulnerabilities, the attack targeted the model's reasoning process itself—highlighting a fundamental shift in how systems can be compromised.

In parallel, recent threat disruption efforts have revealed how adversaries are leveraging AI to orchestrate coordinated influence operations at scale. By chaining multiple AI systems with external platforms and content distribution channels, attackers were able to automate the generation and dissemination of persuasive narratives, significantly increasing both reach and speed. These examples illustrate that AI is not only a target of attack, but also a powerful enabler of adversarial activity. [5, 10]

AI as Target vs AI as Enabler

AI as an Enabler	AI as a Target
✓ Phishing automation	✓ Prompt injection
✓ Deepfake generation	✓ Model manipulation
✓ Malware / exploit development	✓ Sensitive data leakage
✓ Influence / disinformation campaigns	✓ Output exploitation
✓ Automated reconnaissance	✓ Training data poisoning

"AI is shifting from a force multiplier for attackers to a primary attack surface."

Another critical dimension of the threat landscape is the growing exposure of AI systems themselves. As organizations integrate AI into business processes, these systems become directly embedded in decision-making flows and operational environments. This increases the potential impact of compromise, as attacks can influence not only data, but also actions and outcomes.

This shift is further amplified by the expanding ecosystem surrounding AI systems. The use of connectors, plugins, and external integrations allows AI models to access enterprise data, interact with third-party services, and trigger downstream processes. While these capabilities unlock significant business value, they also introduce new trust boundaries and dependencies that are often insufficiently governed, increasing both the likelihood and impact of exploitation.

Taken together, these trends point to a fundamental transformation in the threat landscape. AI is not simply introducing new attack techniques; it is redefining how attacks are executed, scaled, and amplified. The convergence of widespread adoption, increased system interconnectivity, and machine-speed execution is creating an environment where traditional assumptions about detection, response, and control are no longer sufficient.

However, the most profound implication is not just the acceleration of attacks—it is the emergence of AI systems themselves as a primary target. Rather than focusing solely on infrastructure or applications, adversaries are increasingly attempting to manipulate how AI systems interpret inputs, generate outputs, and influence decisions.

The next section explores these emerging risk classes in detail, examining how vulnerabilities in AI systems differ fundamentally from those in traditional software environments. [10, 16]

The New Risk Classes: AI as the Attack Surface

The adoption of AI introduces a fundamentally different category of vulnerabilities compared to traditional software systems. In conventional environments, attackers primarily exploit weaknesses in code, configuration, or infrastructure. In AI-driven systems, however, the attack surface shifts toward how models interpret inputs, learn from data, and generate outputs.

Rather than exploiting code alone, attackers increasingly target **model reasoning, learning processes, and output behaviors**.

This shift transforms AI systems from passive components into **active attack surfaces**, where vulnerabilities emerge dynamically through interaction rather than static flaws. As AI systems become embedded in workflows, decision-making processes, and integrations with external systems, their behavior—not just their code—becomes the primary focus of attack.

Prompt Injection

Prompt injection occurs when attackers manipulate model instructions to override safeguards, alter behavior, or extract sensitive information. Unlike traditional injection attacks, which target execution logic, prompt injection targets **the reasoning layer of the model**.

A notable example is the Varonis "Reprompt" attack, which demonstrated how Microsoft Copilot protections could be bypassed through a coordinated set of techniques. [7, 13]

- **Parameter-to-Prompt (P2P) injection:**
  An attacker embeds malicious instructions into a URL parameter (e.g., the "q" field), which is then passed directly into the model's prompt context. The model interprets this input as legitimate user intent.
- **Double-request technique:**
  The attacker instructs the model to repeat or reinterpret tasks in a way that bypasses initial safety filters. For example, the first response may be benign, but the second request reframes the task to expose restricted information.
- **Chain-request execution:**
  Once the initial interaction is successful, an external server dynamically issues follow-up prompts based on the model's responses, effectively creating a controlled sequence of interactions that guide the model toward unintended behavior.

In practice, this creates a **multi-step reasoning exploit**, where the model becomes part of the attack chain—processing, interpreting, and executing malicious logic without recognizing it as such.

Why traditional controls fail:

- Input validation assumes deterministic parsing, whereas models interpret meaning probabilistically
- Guardrails are applied at individual prompt levels, but attacks unfold across sequences of interactions
- Context windows can be manipulated to reshape instructions and override constraints

Implication:
Prompt injection is not a software flaw—it is a failure to govern how machines interpret and execute instructions.

Insecure Output Handling

AI systems increasingly generate outputs that are directly consumed by downstream systems, workflows, or decision engines. When these outputs are not validated, they introduce a critical risk vector.

Consider the following scenarios:

- **Automation trigger scenario:**
  A generative AI assistant produces a script or command that is automatically executed within an internal system (e.g., infrastructure automation or DevOps pipelines). If manipulated, this output could execute unintended actions such as modifying configurations or exposing data.
- **Business logic manipulation:**
  An AI model used in decision support (e.g., credit approval, fraud detection, or supply chain optimization) produces recommendations that are accepted without verification. An attacker could influence inputs to bias decisions in their favor.
- **Code generation risk:**
  Developers using AI-generated code may unknowingly integrate insecure or malicious logic into production environments if outputs are not reviewed.

In these cases, the AI output is treated as trusted input by downstream systems, effectively bypassing traditional access controls.

Implication:
When AI output becomes executable or decision-driving, it creates a **privilege escalation pathway**, allowing attackers to indirectly influence systems and outcomes without direct access.

Training Data Poisoning

Training data poisoning targets the learning process itself, altering how a model behaves over time. Instead of attacking the system during runtime, attackers compromise the **data used to shape the model's logic**.

Examples include:

- **Embedded trigger scenario:**
  Malicious data is introduced into training datasets, causing the model to behave normally under most conditions but produce specific outputs when triggered by certain inputs.
- **Bias injection:**
  Attackers manipulate datasets to systematically influence model decisions, such as favoring certain outputs or misclassifying specific inputs.
- **RAG pipeline compromise:**
  In retrieval-augmented generation (RAG) systems, an attacker injects malicious content into a knowledge base. When the model retrieves this data, it incorporates it into responses, effectively propagating compromised information.

This transforms training and data pipelines into **attack surfaces**, particularly when organizations rely on external or unverified data sources.

Implication:
Model integrity depends on data integrity—making training pipelines a **critical supply chain risk**.

Model Denial of Service

AI systems are vulnerable to resource-based attacks that exploit their computational and economic characteristics rather than network capacity.

Examples include:

- ✓ **Token flooding:**
  Attackers submit excessively large or complex prompts to increase processing time and resource consumption.
- ✓ **Prompt loops:**
  Carefully designed inputs cause the model to generate recursive or repetitive responses, consuming compute resources.
- ✓ **Context window exhaustion:**
  Attackers fill the model's context window with irrelevant or malicious content, degrading performance or preventing meaningful processing.

In large-scale deployments, these attacks can result in:

- ✓ Service degradation or outages
- ✓ Increased operational costs due to excessive compute usage
- ✓ Disruption of critical workflows reliant on AI services

Implication:
AI denial-of-service attacks introduce both **availability risk and financial exposure**, particularly in consumption-based models.

AI Supply Chain and Plugin Risk

Modern AI systems operate within complex ecosystems of dependencies, including plugins, APIs, retrieval systems, and third-party model providers.

Examples include:

- ✓ **Over-privileged plugin scenario:**
  A plugin integrated into an AI system has access to sensitive data or functionality beyond what is necessary. If exploited, it can expose or manipulate data.
- ✓ **Compromised API dependency:**
  An external API returns manipulated or malicious data, which the AI system incorporates into its outputs.
- ✓ **Untrusted retrieval source:**
  A RAG system retrieves information from an unverified source, introducing malicious or misleading content into responses.

These dependencies create **indirect attack paths**, where attackers exploit integrations rather than the core system.

Implication:
Security must extend beyond the model to govern the **entire AI ecosystem and its dependencies**.

Sensitive Information Disclosure

AI systems may inadvertently expose sensitive information through their outputs due to the way they process and retain context.

Examples include:

- ✓ **Prompt extraction attack:**
  Attackers craft inputs designed to reveal hidden system instructions or internal prompts.
- ✓ **Context leakage:**
  Information from previous interactions is unintentionally included in subsequent outputs.
- ✓ **Inference-based data exposure:**
  Even without direct access, models may reveal sensitive patterns or insights derived from training data.

This is particularly critical when models interact with internal systems, proprietary datasets, or regulated information.

Implication:
Data exposure shifts from direct access to **indirect inference and context leakage**, making it harder to detect and control.

Model Theft and Overreliance

AI models represent valuable intellectual property and can be targeted through extraction or replication techniques.

Examples include:

- ✓ **Model extraction:**
  Attackers repeatedly query a model and analyze outputs to approximate its behavior.
- ✓ **Reverse engineering:**
  Observing patterns in responses to infer underlying logic or training characteristics.

At the same time, organizations face the risk of **overreliance on AI outputs**:

- ✓ Decisions made without human validation
- ✓ Automated processes driven entirely by model output
- ✓ Reduced critical oversight of AI-generated results

Implication:
Overreliance on AI introduces a **decision integrity risk**, where incorrect or manipulated outputs directly impact business outcomes.

Excessive Agency

As AI systems evolve toward more autonomous architectures, they gain the ability to perform actions such as retrieving data, executing tasks, and initiating transactions.

Examples include:

- ✓ AI agents executing workflows across multiple systems
- ✓ Automated financial or operational decisions triggered by model output
- ✓ Multi-step processes performed without human intervention

While this increases efficiency, it also expands the potential impact of compromise. A manipulated system is no longer limited to producing incorrect outputs—it may execute actions at scale.

Implication:
As AI systems gain agency, compromise can lead to **direct operational, financial, and transactional consequences**.

Taken together, these risk classes illustrate a fundamental shift: cybersecurity is no longer only about protecting systems, but about governing how intelligent systems interpret, reason, and act.

As AI systems gain increasing levels of autonomy and integration, traditional control-based approaches become insufficient; and; **as AI systems gain agency, security programs must evolve beyond static defenses toward governance models capable of supervising machine decision-making.**

From AI Risk to Agentic Security

Traditional cybersecurity models were designed to secure systems that execute deterministic logic. These systems behave predictably, operate within defined boundaries, and can be assessed against known failure conditions. As a result, security practices have historically relied on static controls, periodic validation, and predefined response mechanisms to ensure protection.

AI systems fundamentally challenge these assumptions.

Unlike traditional software, AI systems operate through probabilistic reasoning and dynamic context interpretation. Their behavior is not fixed, but continuously shaped by inputs, prompts, data sources, and interactions at runtime. This introduces a level of variability that cannot be fully anticipated or exhaustively tested. Small changes in phrasing, context, or input structure can lead to materially different outcomes, making traditional control validation approaches insufficient.

This shift has significant implications for how security must be approached. In AI-driven environments, risk is no longer static or predictable—it is adaptive, interaction-driven, and capable of evolving in real time. Techniques such as prompt injection and model manipulation illustrate that attackers do not need to exploit code-level vulnerabilities; instead, they can influence how systems interpret, reason, and act. As a result, the attack surface extends beyond infrastructure and applications to include the behavior of intelligent systems themselves.

At the same time, the increasing reliance on connectors, plugins, and external data sources introduces complex interdependencies and blurred trust boundaries. AI systems are embedded within broader ecosystems, interacting with internal systems, third-party services, and external data environments. Each interaction creates a potential pathway for manipulation, making it difficult to define and enforce traditional security perimeters.

Taken together, these dynamics indicate that existing security models—built on static controls, periodic validation, and clearly defined boundaries—are no longer sufficient. This requires a fundamentally new defensive paradigm.

Agentic Security

Agentic security represents an evolution in how organizations design and operate security capabilities in AI-driven environments. Rather than relying solely on predefined controls, agentic security focuses on continuously governing the behavior of intelligent systems as they operate.

In this model, security architectures are designed to:

- ✓ **Continuously sense emerging risks**, monitoring interactions, inputs, and environmental changes in real time
- ✓ **Validate model behavior**, ensuring outputs align with expected policies and constraints
- ✓ **Detect anomalous reasoning patterns**, identifying deviations that may indicate manipulation or misuse

- ✓ **Orchestrate automated containment responses**, enabling rapid mitigation at machine speed

These capabilities reflect a shift from reactive and periodic security toward adaptive and continuous defense.

Several key principles underpin this approach:

- ✓ **Continuous validation rather than periodic assessment**
  Security must operate in real time, validating system behavior as it evolves rather than relying on point-in-time testing
- ✓ **Behavioral monitoring of AI systems**
  Focus shifts from code and infrastructure to how systems interpret inputs and generate outputs
- ✓ **Governance embedded within AI deployment pipelines**
  Security controls must be integrated into the lifecycle of AI systems, from development through deployment and operation
- ✓ **Response capabilities operating at machine speed**
  Detection and response mechanisms must match the speed and scale of AI-driven interactions

Agentic security does not replace existing cybersecurity practices; it extends them to address the unique characteristics of AI systems. It recognizes that in environments where systems can reason and act dynamically, security must evolve from controlling systems to continuously governing their behavior.

This shift also expands the role of security beyond technical control. As AI systems become embedded in decision-making processes and business operations, ensuring their safe and reliable behavior becomes a matter of organizational governance and accountability.

The following section explores how this transformation extends into executive decision-making, requiring new approaches to governance, risk management, and accountability in AI-driven environments.

Governance and Executive Accountability

The rapid adoption of artificial intelligence is expanding the scope of technology risk governance. As AI systems become embedded in core business processes—ranging from customer interaction and decision-making to operational automation—their impact extends beyond technical domains into enterprise-wide risk exposure.

This shift introduces a new set of expectations for boards and executive leadership.

Unlike traditional technology systems, AI-driven capabilities do not simply process data; they interpret, generate, and increasingly influence decisions. As a result, the risks associated with AI are not limited to system failure or data compromise—they include unintended outcomes, behavioral manipulation, and systemic disruptions that can directly affect business performance.

In this context, executive leadership is increasingly required to address a new set of critical questions:

- ✓ **How do we measure exposure to AI-specific risks?**
  Traditional metrics may not capture risks related to model behavior, data dependencies, or interaction-driven vulnerabilities
- ✓ **How resilient are AI-enabled business processes?**
  Organizations must understand how disruptions or manipulations of AI systems impact operational continuity
- ✓ **Can we simulate AI compromise scenarios?**
  Testing must extend beyond infrastructure failure to include scenarios such as model manipulation, adversarial inputs, and automated decision disruption
- ✓ **Do we understand the financial impact of AI system failures?**
  Leaders require visibility into how AI-related risks translate into business impact, including operational disruption, financial loss, and strategic exposure

These questions reflect a broader evolution in governance. AI risk cannot be managed in isolation; it must be integrated into existing enterprise frameworks that guide risk, resilience, and decision-making.

Organizations are therefore increasingly embedding AI-related considerations into:

- ✓ **Enterprise Risk Management (ERM)**
  Expanding risk taxonomies to include AI-specific exposures and ensuring they are measured, monitored, and reported at the enterprise level
- ✓ **Operational Resilience Programs**
  Assessing the ability of AI-enabled processes to withstand, adapt to, and recover from disruption
- ✓ **Technology Governance Frameworks**
  Establishing oversight mechanisms for AI deployment, usage, and lifecycle management

However, integrating AI into governance frameworks is not solely a structural exercise—it requires cross-functional alignment.

AI risk sits at the intersection of multiple domains. Security teams focus on system protection and threat detection; risk management functions assess exposure and impact; legal and compliance teams evaluate regulatory implications; and technology leadership drives implementation and innovation. Effective governance requires these functions to operate in a coordinated manner, with clearly defined roles, shared accountability, and consistent communication.

This also introduces new expectations for transparency and accountability. Executive leadership must be able to demonstrate not only that controls are in place, but that AI systems are being actively governed, monitored, and validated throughout their lifecycle. This includes understanding how models behave, how decisions are influenced, and how risks are identified and mitigated in real time.

Ultimately, governance in AI-driven environments is no longer limited to oversight of systems—it extends to oversight of behavior. Organizations that can effectively integrate AI risk into enterprise governance frameworks will be better positioned to manage uncertainty, maintain resilience, and make informed decisions in increasingly complex environments.

The next section explores how leading organizations translate these governance principles into operational capabilities, turning AI risk into a resilience advantage.

Turning AI Risk into Resilience Advantage

As organizations continue to integrate AI into core business processes, managing risk effectively becomes a source of competitive advantage rather than a constraint. Leading organizations are moving beyond reactive security models and adopting structured approaches that enable them to anticipate, quantify, and respond to AI-driven risks with confidence.

These organizations tend to demonstrate four key capabilities. [1, 6]

AI Resilience Capability Model

Continuous Risk Visibility

AI-driven environments require continuous visibility into risk exposure. Unlike traditional systems, where risks are relatively stable and periodically assessed, AI systems introduce dynamic and interaction-driven risks that evolve over time.

To address this, organizations must continuously monitor:

- ✓ **AI-exposed digital assets**, including models, APIs, and interfaces interacting with users and external systems
- ✓ **Third-party dependencies**, such as external data sources, integrations, and service providers that influence AI behavior
- ✓ **Emerging AI-enabled threat activity**, tracking how adversaries are leveraging AI techniques in real-world scenarios

Continuous visibility allows organizations to identify exposure early, reduce blind spots, and maintain awareness of how AI systems interact with their broader ecosystem.

Risk Quantification and Scenario Analysis

As AI systems become embedded in decision-making processes, executives increasingly require a clear understanding of how technical risks translate into business impact.

This requires moving beyond qualitative assessments toward **quantified risk models** that estimate potential financial and operational consequences. Scenario analysis plays a critical role in this process by enabling organizations to evaluate the impact of different AI-related risk events, including:

- ✓ **AI manipulation**, where adversarial inputs influence system behavior
- ✓ **Model failure**, leading to incorrect or inconsistent outputs
- ✓ **Data leakage**, exposing sensitive or proprietary information

By modeling these scenarios, organizations can better understand potential loss exposure, prioritize investments, and align risk tolerance with business objectives.

Simulation and Preparedness

AI-related incidents introduce new categories of disruption that extend beyond traditional cyber scenarios. These include compromised decision systems, manipulated automated workflows, and large-scale AI service outages.

To prepare for these scenarios, organizations must regularly test their ability to respond through structured simulations and resilience exercises. These simulations should reflect realistic AI-driven incidents and assess:

- ✓ **Decision-making under uncertainty**, where AI outputs may be unreliable or manipulated
- ✓ **Cross-functional coordination**, involving security, risk, legal, and operational teams
- ✓ **Response effectiveness**, including containment, recovery, and communication

Simulation-driven preparedness enables organizations to identify gaps, refine response strategies, and build confidence in their ability to manage complex AI-related disruptions.

Threat Intelligence Integration

Understanding how adversaries are exploiting AI technologies is critical to maintaining an effective defense. AI-enabled threats evolve rapidly, and organizations must continuously adapt their security posture based on real-world attack patterns.

This requires integrating threat intelligence into security operations to:

- ✓ **Track emerging AI-enabled attack techniques**, including prompt manipulation, automated social engineering, and AI-assisted exploitation
- ✓ **Monitor adversary behavior**, identifying how attackers are combining AI capabilities with traditional tactics
- ✓ **Align defensive strategies with observed threats**, ensuring that controls remain relevant and effective

By grounding their defenses in real-world intelligence, organizations can move from reactive security toward proactive risk management.

Taken together, these capabilities enable organizations to shift from managing AI risk as an isolated challenge to embedding it within a broader resilience strategy. Rather than attempting to eliminate risk entirely, leading organizations focus on improving their ability to anticipate, absorb, and adapt to AI-driven disruptions.

This shift transforms AI risk management from a defensive necessity into a strategic enabler. Organizations that can continuously monitor exposure, quantify impact, simulate disruption, and align with evolving threats will be better positioned to operate confidently in environments shaped by intelligent systems.

The final section brings these elements together, outlining how organizations can design for an AI-resilient future.

Conclusion

Designing for an AI-Resilient Future

Artificial intelligence is reshaping both the nature of cyber threats and the capabilities available to defend against them. As organizations continue to integrate AI into critical business processes, the boundary between technology risk and business risk is becoming increasingly blurred.

Organizations that approach AI solely as a productivity or innovation tool risk underestimating its security implications. While AI can drive efficiency, automation, and new forms of value creation, it also introduces fundamentally new categories of risk—driven by probabilistic behavior, dynamic interactions, and interconnected ecosystems. These risks cannot be fully addressed through traditional security models alone.

Resilient organizations are those that recognize this shift early and adapt accordingly.

Rather than focusing exclusively on preventing incidents, they adopt a broader perspective—one that emphasizes governance, continuous oversight, and preparedness. This includes:

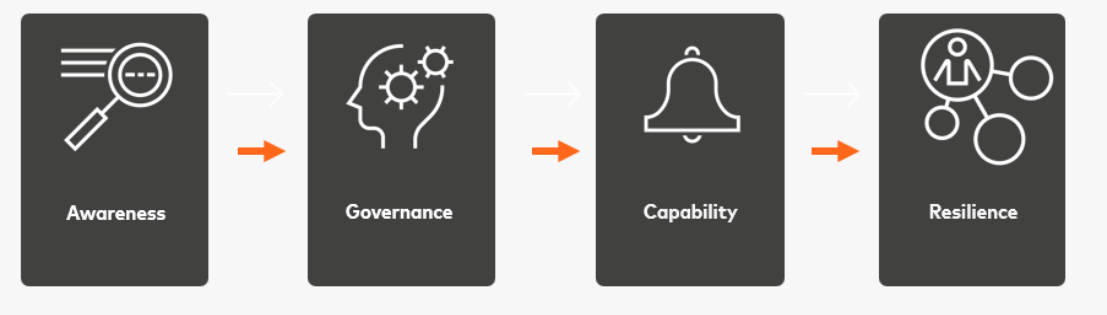
- ✓ **Embedding governance into AI deployment and lifecycle management**, ensuring that accountability and control are integrated from design through operation
- ✓ **Monitoring model behavior continuously**, with a focus on how systems interpret inputs, generate outputs, and influence decisions in real time
- ✓ **Preparing for AI compromise scenarios**, including manipulation of decision systems, disruption of automated workflows, and large-scale service failures
- ✓ **Integrating AI-related risks into enterprise risk frameworks**, enabling leadership to assess, prioritize, and manage exposure in alignment with business objectives

These capabilities reflect a shift from static protection to adaptive resilience.

Ultimately, the next generation of cybersecurity will not be defined solely by stronger defenses or more advanced technologies. It will be defined by the ability of organizations to govern intelligent systems operating at scale—systems that can reason, act, and evolve in ways that traditional models were never designed to address.

In this new environment, resilience is no longer measured by the absence of incidents, but by the ability to anticipate, absorb, and respond to disruption effectively.

AI Resilience Journey



In the next generation of technology risk, resilience will not be measured by how well organizations prevent AI abuse, but by how intelligently they detect, quantify, and contain it. [4, 6, 14]

Learn more

Explore how Mastercard Cybersecurity helps organizations translate emerging AI and technology risk into measurable resilience— supporting visibility, quantification, preparedness, and threat-informed defense across the ecosystem.

Contact us to learn more about Mastercard's Enterprise Cybersecurity portfolio including, Cyber Insights, Cyber Quant, RiskRecon, Cyber Front, Cyber Crisis Exercise and Threat Protection.

References

1. Accenture. (2024). *State of Cybersecurity Resilience 2024*. https://www.accenture.com/us-en/insights/security/state-of-cybersecurity-resilience
2. Cisco. (2026). *State of AI Security Report*. https://www.cisco.com/site/us/en/products/security/state-of-ai-security.html
3. Gartner. (2024). *Top Strategic Technology Trends: AI Governance and Risk Management*. https://www.gartner.com/en/articles/top-strategic-technology-trends
4. ISO/IEC. (2023). *ISO/IEC 42001: Artificial Intelligence Management System (AIMS)*. https://www.iso.org/standard/81230.html
5. KELA. (2025). *AI Threat Report: How Cybercriminals are Weaponizing AI Technology*. https://www.kelacyber.com/resources/research/2025-ai-threat-report/
6. McKinsey & Company. (2024). *The State of AI: Global Survey*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai
7. Microsoft. (2024). *Defending Against AI-Enabled Threats*. https://www.microsoft.com/en-us/security/blog/2024/05/08/defending-against-ai-enabled-threats/
8. MITRE ATT&CK Framework
9. Netskope. (2026). *Cloud and Threat Report 2026*. https://www.netskope.com/netskope-threat-labs/cloud-threat-report-2026
10. OpenAI. (2026). *Disrupting Malicious Uses of AI*. https://openai.com/index/disrupting-malicious-ai-uses/
11. OWASP. (2023). *Top 10 for Large Language Model Applications*. https://owasp.org/www-project-top-10-for-large-language-model-applications/
12. Palo Alto Networks Unit 42. (2025). Global Incident Response Report 2025. https://www.paloaltonetworks.com/resources/research/unit-42-incident-response-report-2025
13. Varonis Threat Labs. (2026). *Reprompt: The Single-Click Microsoft Copilot Attack that Silently Steals Your Personal Data*. https://www.varonis.com/blog/reprompt
14. World Economic Forum. (2024). *Global Cybersecurity Outlook 2024*. https://www.weforum.org/publications/global-cybersecurity-outlook-2024/
15. Zscaler. (2026). *ThreatLabz 2026 AI Security Report*. https://www.zscaler.com/press/zscaler-2026-ai-threat-report-83-year-over-year-surge-ai-activity-creates-growing-oversight
16. Europol. (2024). *Internet Organised Crime Threat Assessment.* https://op.europa.eu/en/publication-detail/-/publication/f7e2258d-4b04-11ef-acbc-01aa75ed71a1/language-en