Securing AI: Six steps to enable trusted innovation while addressing risk

A roadmap to integrate AI security into enterprise operations, from initial discovery to continuous validation.

kpmg.com

Image created with aIQ Gemini

The next evolution of leading-edge cybersecurity

Image created with aIQ Gemini

Everyone is talking about AI security—the practice of extending traditional cybersecurity to safeguard AI systems by protecting data, models, and actions from emerging risks like bias, tampering, and adversarial attacks. Across enterprises, business units are already leveraging Agentic AI and GenAI tools to accelerate innovation and accomplish low-level tasks. Yet too often, application and infrastructure security teams aren't included early enough to ensure these enhancements and new AI use cases are secure and effective, which leaves organizations exposed to growing AI risks.

In the rush to adopt AI, many enterprises have focused on rapid enablement and expanding capabilities, hoping their security programs can keep up. Spoiler alert: They can't. Pilot projects are outpacing protection, scattering accountability, and widening the gap between what AI can do and what organizations can control.

Rather than rebuilding security programs from scratch, organizations are working to apply an AI lens to their existing frameworks—often retroactively—because many AI deployments bypassed typical security steps and controls. Others are realizing that their current programs can be extended and adapted with the right updates and consistency. At a minimum, every organization should establish clear policies and governance before any AI tools are leveraged to avoid building on an inefficient, ineffective, and unstable foundation that undermines long-term security and trust.

The fundamental challenge is that AI creates new dependencies and responsibilities that traditional controls weren't designed to manage. It significantly expands the attack surface, exposing new vulnerabilities like bias, hallucination, model tampering, social influence engines, and unpredictable decision-making. These systems can also act autonomously on behalf of users, often with elevated privileges—making strong authentication and authorization essential to prevent abuse. Security frameworks must now account for systems that can act autonomously and steadily change how they operate. Securing these systems isn't just about new policies and tools—it demands a disciplined, programmatic approach that can keep pace with AI's rapid evolution.

82% of leaders in a recent KPMG CEO survey cited cybersecurity as their company's top threat amid AI's growing risks.

And executing on that mission tops the CEO's agenda as well: 82 percent of leaders in a recent KPMG CEO survey cited cybersecurity as their company's top threat amid AI's growing risks. For security teams, the way forward is to evolve existing programs—strengthening what already works while layering on the governance, validation, and continuous monitoring needed to manage the many "we haven't seen this before" challenges of AI.

Here are six practical steps to make that happen. It's an iterative roadmap to help organizations move from high-level discussion to tactical execution and establish a scalable, trusted AI security program built on leading strategies, relentless innovation, and an enterprise-wide commitment to trust.

Define your AI security strategy

Before any AI or technology initiatives begin, the chief information security officer (CISO) and team, working in partnership with business leaders across the enterprise, must fully understand the organization's overall AI and GenAI language models plans and goals—and where security fits within that strategy. AI is moving faster than most security programs can adapt, and this capability gap means that even early engagement may not guarantee effective governance.

Effective AI risk management starts with coordinated governance and clear accountability. Security leaders (including CISOs), technology teams (such as CIOs), and business stakeholders should embed visibility and controls into every AI initiative—not as an afterthought, but as part of the design. Large language models need to be integrated into risk frameworks early, with ownership clearly defined and security aligned to enterprise goals. When strategy and security evolve together, organizations scale AI faster and with fewer missteps.

Establishing a clear, AI-specific security strategy from the start brings alignment quickly. It defines ownership, sets measurable goals for secure adoption, and matches enterprise ambition with informed governance. When strategy and security evolve together, AI initiatives scale faster, and with fewer missteps. In this way, security leaders move from gatekeepers to enablers of innovation, embedding trust, compliance, and durability into every AI decision from Day 1, rather than bolting them on later.

Making it happen

- **Clarify enterprise AI objectives:** Partner with business and data leaders to pinpoint where AI will create value (e.g., operations, customer engagement, finance, etc.) and understand the intended data, technology, and resources to be leveraged.
- **Build cross-functional alignment:** Establish a working group of stakeholders from security, data, compliance, legal, and key business units to coordinate policy updates and communicate risk priorities to leadership.
- **Define the AI security mandate:** Translate enterprise goals into a strategy for securing AI and document accountability through formal AI security responsibilities across domains such as data protection, access governance, model assurance, and continuous monitoring. Include documentation of escalation paths for AI-related risk.
- **Set measurable outcomes:** Determine how success will be tracked—for example, a steady reduction in unmanaged use cases and faster validation cycles—and align metrics with enterprise key performance indicators (KPIs).
- **Integrate cyber risks into the corporate risk register:** Formally document and track cybersecurity and AI-related exposures alongside other enterprise risks, closing gaps that are too often missed in traditional risk management.

Know where you are

You can't secure what you don't know. Once the overall security strategy is in place, visibility becomes the next big priority. Many organizations advance quickly in AI experimentation but lag in establishing guardrails and control. Most are still in the early phases of maturity—experimenting rather than governing. Real progress comes when security is built in from the start by secure-by-design practices, testing and validation, and continuous runtime monitoring of these solutions throughout the AI lifecycle, rather than being treated as an afterthought.

A critical goal is to create a single, trusted view of the enterprise's AI landscape: what systems exist, who owns them, how they operate, and where the potential risks lie. This requires identifying and defining everything that qualifies as AI—including systems, processes, and agentic workflows that are often overlooked—and validating which assets truly include AI components. That comprehensive visibility forms the foundation for every control, validation, and monitoring decision that follows. Mature AI security programs treat this attention to detail as an ongoing lifecycle, from discovery through validation and monitoring, ensuring that both visibility and assurance grow together over time.

Making it happen

- **Run an AI maturity diagnostic:** Begin with a structured assessment across technology, governance, and people. Benchmark against NIST frameworks or other leading standards to identify strengths and close critical gaps.
- **Map your AI footprint:** Inventory all models, datasets, and third-party integrations in use, whether established or experimental. Identify and include "shadow AI" tools that may be operating outside formal oversight. Attributes of this inventory are often referred to as an "AI Bill of Materials" to showcase all the components that make up an AI system, including additional context such as ownership, dependencies, and intended use case / audience. Use this to support compliance reporting and ongoing validation.
- **Tier your risks:** Develop a risk assessment that covers the cross-functional domains as well as cyber. Leveraging this scoring, document the risk score of the system informed by overarching business criticality, data sensitivity, and exposure level. Apply those tiers to prioritize testing, controls, and monitoring.

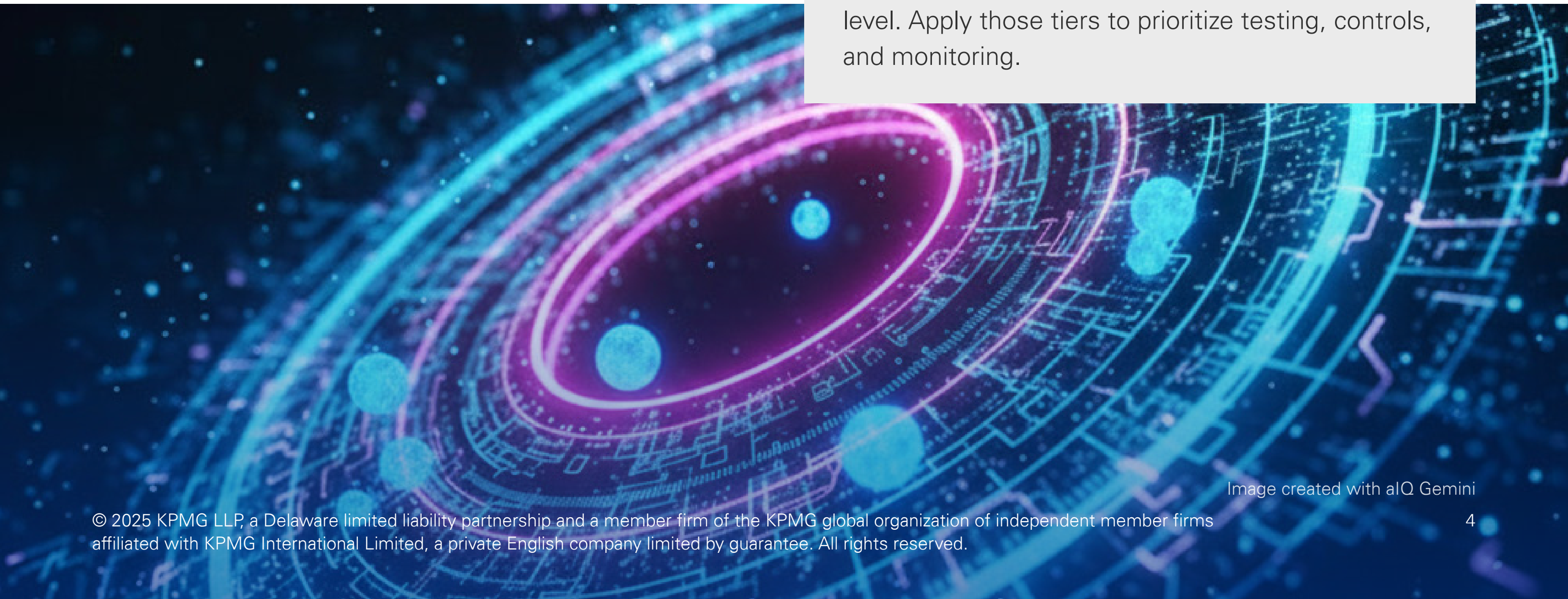


Image created with aIQ Gemini

Strengthen your security framework for AI

AI security is the next evolution of cybersecurity, not a separate discipline. Rather than rebuilding from scratch, leading organizations are expanding their existing programs to account for AI's materially different risk profile. That means updating core domains—identity and access management, data protection, and application security—to accommodate automated business processes, AI-related data flows, and decision logic.

The objective is to strengthen the foundation already in place by aligning established cyber practices with the dynamic behavior of AI systems. When done well, this integration helps AI initiatives scale securely, with consistency, transparency, and control.

Making it happen

- **Detail AI's impact on existing domains:** Review the major cybersecurity arenas through an AI lens: identity, access, privacy, data protection, application security, incident response, and other relevant domains. And then determine AI's impact and where processes must evolve to handle things like securing the usage of MCP servers, standardized logging approaches for agents invoicing tools, and establishing observability across AI systems and agent calls.
- **Integrate AI into governance routines:** Embed AI-related risk discussions into standing security councils and change-management groups. Require that all AI use cases follow the same intake, approval, and documentation processes as any other critical technology.
- **Extend existing frameworks:** Build on what already works. For example, if your organization follows the NIST Risk Management Framework (RMF), align your existing RMF controls and processes with the emerging NIST AI RMF—mapping current security and privacy safeguards to new AI risk areas such as data quality, model transparency, and accountability.
- **Reinforce accountability:** Update policies and job roles so that ownership of AI systems is explicit—from model development and deployment to continuous validation and monitoring.
- **Automate for scale:** As AI adoption grows, introduce automation through AI TRiSM (trust, risk, and security management) or discovery tools to streamline oversight, detect policy violations, and flag unapproved model usage.

Build and integrate effective controls

With core security domains updated, the next priority is embedding targeted AI controls. A unified control framework that spans security, privacy, and compliance creates the guardrails that make AI innovation safe and defensible. Controls must fit seamlessly into existing processes, evolving with models and regulations, while remaining measurable and auditable.

Organizations that embed AI controls into their established cyber and risk frameworks gain consistency, accelerate validation, and maintain confidence that their overall approach continues to work as intended, even as systems and threats evolve.

Making it happen

- **Map AI risks to recognized frameworks:** Align program design with frameworks and standards such as the NIST AI RMF, ISO/IEC 42001, and the EU AI Act. This ensures controls address regulatory expectations while staying consistent with enterprise risk strategies
- **Define clear control categories:** Focus on key areas (e.g., model integrity, data provenance, access management, output validation, auditability, business ownership) and specify how each will be monitored and reported.
- **Evaluate control effectiveness and resilience:** When defining and evaluating AI controls, ensure there are clear timelines for how to move from design to fully operational controls. Include expectations for incident response, business continuity, and disaster recovery planning so the business can maintain operations or quickly resume in the event of a control failure, minimizing disruption and prioritizing resilience.
- **Avoid the parallel-governance trap:** Embed AI control checks directly into existing change-management workflows, risk registers, and assurance testing instead of creating a separate process.
- **Require validation before release:** Make independent review and formal attestation of a standard gate before any AI system and agentic workflow goes live, supported by documented testing and sign-off confirming controls are effective.

Conduct rigorous validation and testing

Controls are only as effective as the testing behind them. Validation turns governance from a checklist into a living practice, demonstrating that controls work, risks are contained, and AI systems behave as intended. Testing must be systematic, repeatable, and continuous across the model lifecycle.

The goal is to confirm that every AI system, from development to deployment, meets its defined security and compliance standards and that vulnerabilities are caught early rather than after launch.

Image created with aIQ Gemini

Making it happen

- **Build testing into development:** Treat validation as part of the build process, not a final step. Integrate AI testing checkpoints into continuous integration and deployment (CI/CD) pipelines to catch issues before release.
- **Apply the right depth of scrutiny:** Tailor validation to each system's risk tier—models with higher business impact or sensitive data exposure require deeper, more frequent testing.
- **Use multiple testing methods:** Combine static and dynamic testing (SAST, DAST, and SCA) with AI-specific techniques such as adversarial/red-teaming (inclusive of prompt injection, and data-poisoning simulations).
- **Test your response readiness:** Run an annual table-top exercise with business executives to ensure there are no gaps in understanding response protocols should they need to be enacted.
- **Document and attest:** Produce formal records of testing results through AI System Cards or equivalent reports. These attestations build traceability and support both internal assurance and regulatory readiness.
- **Close the loop on findings:** Create a defined feedback process so vulnerabilities identified in validation feed directly into risk remediation, model retraining, or control enhancement.

Step 0 Step 1 Step 2 Step 3 Step 4 **Step 5**

Ensure continuous monitoring and adaptive security

AI models evolve constantly—and so do the threats that target them. Once controls and validation are in place, the focus shifts to comprehensive, continuous oversight. Monitoring confirms that systems stay within approved risk thresholds as they learn, retrain, and interact with new data. The objective is to move from periodic checks to real-time visibility, using automation and AI-driven analytics to detect anomalies early, respond quickly, and sustain assurance as the environment changes.

Making it happen

- **Establish runtime monitoring:** Track model drift, data exfiltration, and performance anomalies in real time. Integrate automated alerts with security operations for unified incident response.
- **Correlate AI signals with enterprise risk:** Feed AI-specific telemetry—access patterns, model outputs, training-data changes—into enterprise risk dashboards to connect technical activity with business impact.
- **Automate adaptive response:** Use machine learning and workflow automation to reevaluate controls dynamically and retrain models when thresholds are breached. This keeps security posture current without manual intervention.
- **Refine through threat intelligence:** Integrate insights on emerging AI attack techniques into the monitoring environment to anticipate and mitigate new risks before they escalate.
- **Assess and retrain the workforce:** Test employees regularly on AI security protocols and retrain as needed to ensure awareness and readiness keep pace with evolving threats.
- **Get more help:** Extend monitoring capacity through managed detection and response or other 24/7 assurance services. These models combine human expertise with automated analytics to maintain continuous protection at scale.

Image created with aIQ Gemini

Building a culture of trusted AI

AI security doesn't belong to one team. It's an enterprise responsibility that depends on shared trust, transparency, and accountability. Every business function has a role in governing how AI is designed, deployed, and refined. But turning that principle into practice requires clear leadership and integration.

That's where the CISO comes in. Their mission now extends beyond protecting systems to architecting how AI operates safely across the organization. The CISO connects the technical, ethical, and regulatory threads, aligning cyber, data, and compliance teams so every new AI use case enters a controlled, measurable environment. This doesn't require owning every decision but ensures each decision falls within consistent, enforceable boundaries that expand as adoption grows.

From there, the mandate becomes execution. Securing AI is the next evolution of the core cybersecurity mission: visibility, validation, and accountability at speed. The best programs build frameworks that can absorb change, automate assurance, and learn as fast as the models they protect. That's how CISOs can lead both innovation and protection at the same time—and move trusted AI from aspiration to reality.

How KPMG can help

Implementing an AI security program requires the tools, talent, and testing capacity to sustain it. KPMG Cyber Managed Services can help organizations operationalize trusted AI through a flexible mix of consulting engagements and managed offerings, including managed, comanaged, and advisory support.

Our services include:

AI security assessment and readiness: Evaluate AI maturity, identify governance and control gaps, and benchmark against leading frameworks such as the NIST AI RMF.

Develop IAM capabilities for AI: Assess and uplift your existing IAM governance structures to ensure coverage for AI systems and NHI (non-human identities) for AI agents.

Managed testing and validation: Conduct continuous red-teaming, adversarial testing, and model validation through the KPMG Cyber Managed Testing platform—combining automation, threat intelligence, and expert oversight.

Continuous monitoring and response: Deliver 24/7 visibility into model drift, data exfiltration, and policy violations, integrated with your existing SOC operations.

Governance integration and automation: Embed AI risk controls and reporting into enterprise risk management systems and automate assurance through AI TRiSM and workflow orchestration tools.

Program management and optimization: Provide ongoing support for roadmap execution, control attestation, and regulatory readiness as AI systems scale.

The KPMG Managed Services model enables clients to add specialized AI security support quickly, control costs, and maintain continuous assurance without expanding headcount. Please reach out to our team to learn more.

Meet our team

Our AI security and cybersecurity managed services professionals combine deep technical skills with real-world experience to help organizations secure their enterprise systems. Our teams include specialists in AI assurance frameworks, model validation, and risk governance—working alongside other specialists in threat management, identity, compliance, and other core cyber domains to deliver the required protection. Using a human-led, technology-enabled approach, we integrate automation and continuous monitoring to strengthen defenses and accelerate response. Every engagement is designed to scale with clients' needs—bringing the right mix of insight, execution, and managed services to keep innovation secure and compliant.

Katie Boswell

Managing Director

Cyber Security & Tech Risk Management

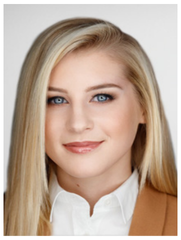
katieboswell@kpmg.com

Chris Crevits

Principal, Advisory

Cyber Managed Services

ccrevits@kpmg.com



Kristy Hornland

Director Advisory

Cyber Security Services

khornland@kpmg.com

Weston Cole

Specialist Director

Delivery Management

westoncole@kpmg.com

Griffin Reid

Specialist Director

Cyber Managed Services

griffinreid@kpmg.com

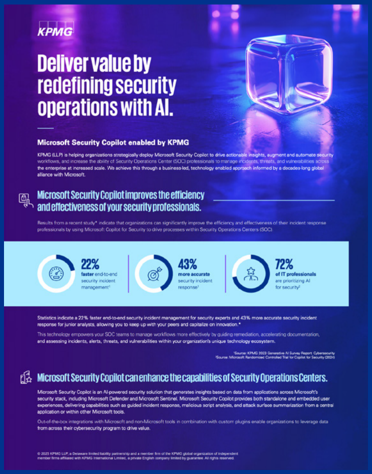
Sailesh Gadia

Partner, Advisory

Cyber Security Services

sgadia@kpmg.com

Related thought leadership:



KPMG

Deliver value by redefining security operations with AI.

Microsoft Security Copilot enabled by KPMG

KPMG (LLP) is helping organizations strategically deploy Microsoft Security Copilot to drive actionable insights, augment and automate security workflows, and increase the ability of Security Operations Center (SOC) professionals to manage incidents, threats, and vulnerabilities across the enterprise at increased scale. We achieve this through a business-led, technology enabled approach informed by a decades-long global alliance with Microsoft.

Microsoft Security Copilot improves the efficiency and effectiveness of your security professionals.

Results from a recent study* indicate that organizations can significantly improve the efficiency and effectiveness of their incident response professionals by using Microsoft Copilot for Security to drive processes within Security Operations Centers (SOC).

22% faster end-to-end security incident management[1]

43% more accurate security incident response[1]

72% of IT professionals are prioritizing AI for security[2]

Statistics indicate a 22% faster end-to-end security incident management for security experts and 43% more accurate security incident response for junior analysts, allowing you to keep up with your peers and capitalize on innovation.*

This technology empowers your SOC teams to manage workflows more effectively by guiding remediation, accelerating documentation, and assessing incidents, alerts, threats, and vulnerabilities within your organization's unique technology ecosystem.

[1]Source: KPMG 2023 Generative AI Survey Report: Cybersecurity
[2]Source: Microsoft Randomized Controlled Trial for Copilot for Security (2024)

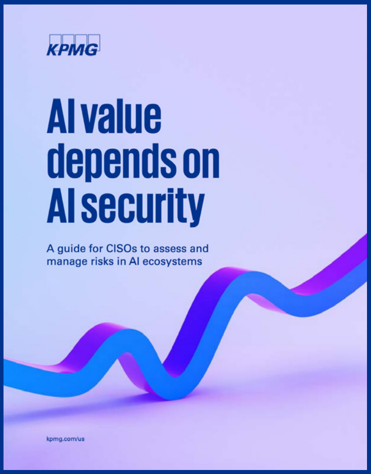
Microsoft Security Copilot can enhance the capabilities of Security Operations Centers.

Microsoft Security Copilot is an AI-powered security solution that generates insights based on data from applications across Microsoft's security stack, including Microsoft Defender and Microsoft Sentinel. Microsoft Security Copilot provides both standalone and embedded user experiences, delivering capabilities such as guided incident response, malicious script analysis, and attack surface summarization from a central application or within other Microsoft tools.

Out-of-the-box integrations with Microsoft and non-Microsoft tools in combination with custom plugins enable organizations to leverage data from across their cybersecurity program to drive value.

© 2025 KPMG LLP, a Delaware limited liability partnership and a member firm of the KPMG global organization of independent member firms affiliated with KPMG International Limited, a private English company limited by guarantee. All rights reserved.

Delivery value by redefining security operations with AI



KPMG

AI value depends on AI security

A guide for CISOs to assess and manage risks in AI ecosystems

kpmg.com/us

AI value depends on AI security

Advisory | AI Governance > | Cyber Security >

Unleashing the Power of AI: the KPMG Pioneering Approach to AI Security

Artificial intelligence (AI) has emerged as a transformative force across industries, reshaping business processes, enhancing decision-making, and driving unprecedented efficiencies. However, as AI systems grow in complexity and capability, regulating their reliability, security, and fairness becomes paramount. Organizations must elevate their teams as well as the tools enabling them to drive innovation that can endure evolving risks and regulations.

Unleashing the Power of AI: the KPMG Pioneering Approach to AI Security

Beyond 'managed' security: 5 ways to close the cyber performance gap

Can your cybersecurity keep pace with AI-powered threats? Discover how future-ready cyber services are meeting the challenge.

Cyber threats are moving faster than most enterprises can respond

Cyber threats are evolving faster than enterprise defenses. Adversaries now leverage AI-driven reconnaissance, automated exploits, and machine-

Beyond 'managed' security: 5 ways to close the cyber performance gap

Some or all of the services described herein may not be permissible for KPMG audit clients and their affiliates or related entities.

Please visit us: in | **kpmg.com** | **Subscribe**

The information contained herein is of a general nature and is not intended to address the circumstances of any particular individual or entity. Although we endeavor to provide accurate and timely information, there can be no guarantee that such information is accurate as of the date it is received or that it will continue to be accurate in the future. No one should act upon such information without appropriate professional advice after a thorough examination of the particular situation.