

Don't Leave AI Exposed

| A guide on how to protect against
| the top AI challenges



Redefining cybersecurity: Cloud, AI and the new imperative for trust

AI is accelerating innovation across every industry, transforming how organisations design, deploy and protect new digital capabilities. From predictive maintenance in manufacturing to personalised learning in education, AI apps and agents are reshaping business functions with unprecedented speed and intelligence. But as these intelligent systems proliferate across sectors – from healthcare diagnostics to financial fraud detection – the complexity and scale of digital environments are growing exponentially.

Traditional security models can't keep up. A unified approach – one that connects visibility, automation and governance across both cloud and AI – is now essential to maintaining trust and momentum.

This guide explores what that approach looks like in practice. It examines how cloud security is being reshaped by AI, new threats emerging from intelligent systems and how security leaders and developers can strengthen defences, protect sensitive data and stay ahead of evolving threats and compliance requirements – accelerating business transformation while enhancing security.

Key insights you'll gain:

- ✔ **AI is scaling fast – and so are the threats that follow.** How rapid AI adoption is reshaping cloud security and stretching visibility, operations and governance.
- ✔ **How AI is rewriting the rules of cyber defence.** The new attack techniques targeting data, models and systems – and what they mean for defenders.
- ✔ **The four challenges standing between your organisation and secure AI adoption.** Guidance for security leaders and developers to strengthen visibility, protect data and stay compliant.
- ✔ **Where proven security meets the new world of AI.** How Microsoft Defender for Cloud and Microsoft Purview extend unified protection across environments.
- ✔ **Securing the future of AI innovation together.** Next steps and resources to continue strengthening AI security.

AI is scaling fast – and so are the threats that follow

While the value of AI-enabled innovation is undeniable, it also introduces new levels of complexity in cloud security and creates new openings for attackers.

68%

of security leaders cite inadequate visibility and monitoring across multicloud and hybrid environments as their top concern.¹

AI is making cloud security more complex

The same scalability that enables AI also increases complexity. Each new model, dataset or API connection adds potential vulnerabilities and expands the attack surface. Fragmented tools and inconsistent policies leave defenders without the context needed to assess risk across workloads.

89%

of leaders cited a YOY increase in cloud security incidents to an average of nine per year.²

Growing pressure on security teams

AI workloads introduce new pipelines, orchestrators and third-party integrations that stretch security operations. Sensitive information can appear across training data, models and outputs, while alert volumes rise across hybrid environments.

55%

of leaders lack understanding of how AI is regulated.³

Regulatory demands are exploding

Regulatory frameworks such as the EU AI Act and NIST AI Risk Management Framework (RMF) are redefining expectations for accountability and compliance. With 250+ new regulations each day – and continued uncertainty around the regulation of AI itself – organisations need to streamline and automate governance.

In addition to existing threat vectors, AI introduces new attack surfaces, risks and threats. We'll examine these emerging challenges in more detail in the next chapter.

AI is rewriting the rules of cyber defence

Beyond complicating existing threat vectors, AI introduces new risks based on how it processes and connects data at scale. Its ability to interpret prompts, generate responses and draw from

retrieval-augmented data expands the attack surface. As models link with AI orchestration layers and external plugins, new openings emerge that traditional defences were not built to protect.



Prompt injection

Attackers craft malicious prompts or jailbreak instructions that cause models to ignore safeguards and execute unintended commands. These prompts often appear harmless, making them difficult to detect through rule-based controls.

Gartner research reports that 85% of organisations are concerned about indirect prompt injection attacks.⁴

Data leakage and shadow AI

Employees using unmanaged AI tools can expose sensitive information to unauthorised parties inside or outside the organisation, including attackers. Even approved AI systems may surface confidential content during training or inference without proper safeguards in place.

80% of leaders cited leakage of sensitive data as their primary concern.⁵

Model abuse and life cycle threats

Adversaries manipulate training data – through mislabelled or tampered inputs – to alter a model's performance. Corrupted models may operate normally while embedding bias or error into outputs. Detecting these manipulations requires continuous validation and dataset lineage tracking before and after deployment.

Gartner research reports that 85% of organisations are concerned about indirect prompt injection attacks.⁴

Evasion attacks

Attackers subtly modify inputs to deceive models into misclassifying or ignoring threats. These exploits bypass filters by targeting model behaviour, making adaptive, AI-powered monitoring essential for detection.

AI is expanding the range of threats, from subtle prompt attacks to complex data manipulation. The next chapter examines how unified protection across cloud and AI environments helps defenders stay ahead of these evolving risks.

From risk to resilience: Tackling AI's top security challenges

AI is changing not just what needs to be protected, but how protection works. Its interconnected data flows, dynamic models and autonomous agents require defences that adapt continuously rather than react

to fixed threats. Securing this environment means extending proven practices – visibility, governance and automation – into an integrated system built for constant learning and response.

Challenge:

The proliferation of AI pushes the security perimeter into new, constantly shifting territory.

Guidance:

For security leaders: Inventory all cloud and AI assets, including agents and models, to strengthen visibility across environments. Establish governance boundaries and access management, including assigning an identity to each agent, to maintain control of expanding identities and workloads.

For developers: Define and implement life cycle policies for AI agents in code and workflows, ensuring their identities, permissions and behaviours are continuously validated, from code to runtime.

Tools to support: Unified identity context, telemetry and configuration analysis reduce exposure and maintain control. Automated monitoring and incident-response plans address AI-specific threats, while ongoing tracing and alerts detect drift and maintain reliability in production.

Challenge:

Sensitive data now moves through complex AI pipelines and model interactions, heightening the risk of exposure.

Guidance:

For security leaders: Apply labelling and access boundaries to protect data throughout the AI life cycle and prevent oversharing. Automate discovery and classification to maintain visibility into sensitive data across clouds and models.

For developers: Build and deploy models with secure-by-default data handling.

Tools to support: Data security posture management and adaptive access controls extend protection wherever data travels. Automated discovery and classification maintain visibility into sensitive information, while dynamic policies help prevent oversharing, block unauthorised data sharing in real time and ensure consistent safeguards across cloud and AI environments.

Challenge:

Emerging attack techniques – from prompt injection to model manipulation – are bypassing static defences and exploiting unseen vulnerabilities.

Guidance:

For security leaders: Use continuous AI risk assessment and adaptive analytics to detect and prioritise evolving threats.

For developers: Integrate vulnerability testing and detection into the DevSecOps workflows by simulating adversarial attacks.

Tools to support: Threat intelligence, runtime protection and behavioural analytics provide proactive insight into AI-specific risks. Red teaming identifies and fixes vulnerabilities before and after deployment, improving model and agent resilience against adversarial attacks.

Challenge:

Rapidly evolving regulations require consistent oversight of AI systems and data.

Guidance:

For security leaders: Work with compliance leads to align AI systems with regulations and standards such as the EU AI Act and NIST AI RMF, establish governance frameworks early in the AI life cycle and automate compliance monitoring.

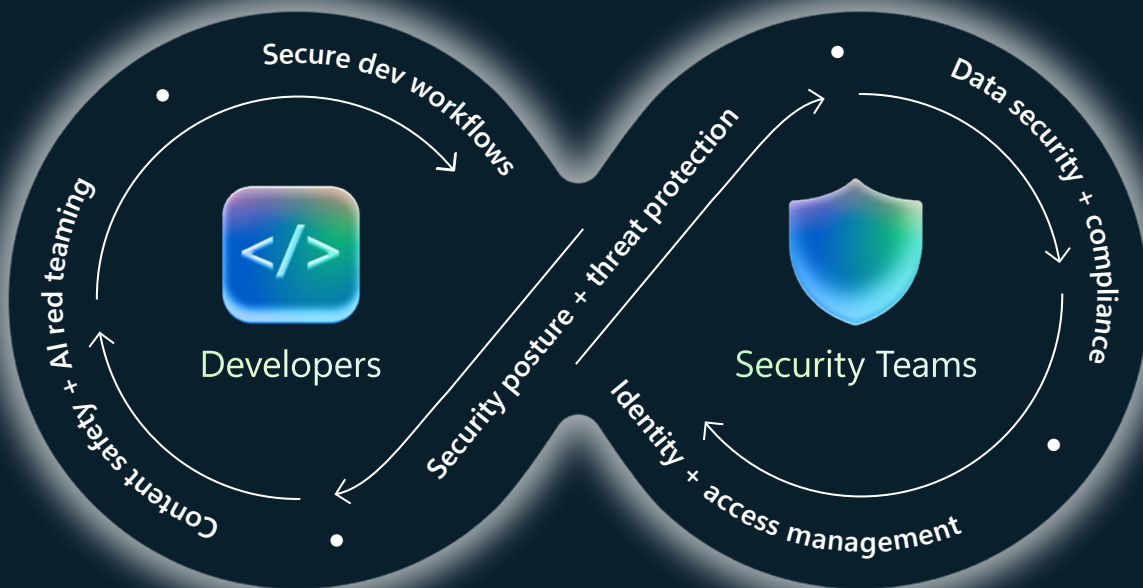
For developers: Build with explainability and data governance in mind, ensuring every model decision can be traced for compliance and audit reviews.

Tools to support: Integrated governance frameworks and automated evidence collection keep oversight current as requirements change. Continuous compliance assessments track regulatory updates, while centralised reporting simplifies audits and demonstrates accountability across cloud and AI systems.



Where trusted security meets the new world of AI

Protect cloud and AI from code to runtime



Start secure and stay secure with Microsoft



Microsoft Defender
for Cloud

Continuous threat
detection & protection



Microsoft
Purview

Data security &
governance



Microsoft
Foundry

Responsible AI &
built-in content safety

Unified protection for cloud and AI



Unify visibility and control across environments

Defender for Cloud, Purview and Microsoft Foundry bring together context from workloads, APIs and data to identify misconfigurations and vulnerabilities – strengthening posture management and helping teams secure assets.



Prevent data exposure and strengthen protection

Integrated detection and response correlate signals from cloud workloads and AI apps and agents to identify prompt injection, data leakage and abnormal activity – containing threats quickly and preserving privacy in sensitive environments.



Defend against evolving AI threats

Continuous monitoring and Microsoft Threat Intelligence detect new attack patterns across hybrid and multicloud systems, using AI-powered analytics to surface vulnerabilities and prioritise response.



Stay compliant and build trust

Defender for Cloud and Purview centralise and automate compliance assessments with 450+ built-in assessments and templates covering 300+ global regulations and standards, such as the EU AI Act and NIST AI RMF, to help organisations maintain transparency and meet evolving regulatory demands.

Hear from those we protect

KPMG: Purview DSPM for AI unifies AI visibility

"DSPM for AI is the reporting console where we see all our policies and controls, along with how data across the organisation has been classified. [...] It's a one-stop shop portal to report on detection of business risks and regulatory violations."

Greg Schellenberg, Director of Cyber Security, KPMG

[Read the full story →](#)

Mia Labs: Defender for Cloud blocks jailbreaks

"We've successfully resisted multiple jailbreak attempts. Defender for Cloud shows us how to design our processes with optimal security and monitor where jailbreak attempts may have originated."

Marwan Kodeih, Chief Product Officer, Mia Labs, Inc.

[Read the full story →](#)

EY: Purview SDK reduces development time up to 30%

"We're not just building AI tools, we are creating agentic solutions where trust, security and transparency are present from the start, supported by the policy controls provided through the Purview SDK."

Sumanta Kar, Partner, Innovation and Emerging Tech, EY

[Read the full story →](#)

Securing the future of AI innovation together

AI is transforming how organisations protect data, workloads and applications. The same intelligence that accelerates innovation now strengthens defence, connecting insights across cloud and AI environments to detect risks, automate response and maintain compliance at scale.

For IT leaders, it means unified visibility and control. For developers, it means secure systems and workflows that keep pace with innovation. Together, these capabilities help organisations build and deploy AI securely – with protection that evolves as fast as progress.



Download [the datasheet](#)



Explore [Security for AI](#) solutions



Learn more about [Cloud Security](#)



© 2025 Microsoft Corporation. All rights reserved. This document is provided 'as-is'. Information and views expressed in this document, including URL and other internet website references, may change without notice. You bear the risk of using it. This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal reference purposes.

¹ IDC White Paper, Sponsored by Microsoft, ['THE NEXT ERA OF CLOUD SECURITY: Cloud-Native Application Protection Platform and Beyond'](#), Doc. #US53297125, April 2025

² IDC White Paper, Sponsored by Microsoft, ['THE NEXT ERA OF CLOUD SECURITY: Cloud-Native Application Protection Platform and Beyond'](#), Doc. #US53297125, April 2025

³ 'First Annual Generative AI Study: Business Rewards vs. Security Risks', iSMG

⁴ Gartner®, Gartner Peer Community Poll – If your org's using any virtual assistants with AI capabilities, are you concerned about indirect prompt injection attacks? (Aug 2025) <https://www.gartner.com/peer-community/poll/if-orgs-using-virtual-assistants-ai-capabilities-concerned-about-indirect-prompt-injection-attacks>

⁵ First Annual Generative AI Study: Business Rewards vs. Security Risks', iSMG

⁶ Gartner Predicts 40% of AI Data Breaches Will Arise from Cross-Border GenAI Misuse by 2027, Gartner