Microsoft Security

Issue 01

Microsoft Guide for Securing the AI-Powered Enterprise

Getting Started with AI Applications

> According to Microsoft research, 47% of users are "very confident" that generative AI security solutions can make critical security decisions in high-stakes scenarios.

The briefing

Getting started with AI applications

Artificial intelligence (AI) is changing the way we work and reshaping how businesses operate. Chances are that you and your organization have already begun exploring or are just starting to use AI applications—but alongside these opportunities come new risks that demand immediate attention. Employees may also be adopting consumer-grade AI tools on their own (often referred to as shadow AI), creating additional challenges for security and oversight.

AI applications thrive on data, can integrate deeply into workflows, and often make critical decisions in real time. Without the right security measures, they can expose sensitive information, introduce vulnerabilities, and create compliance challenges. As organizations embrace AI, they must also navigate evolving regulations like the **EU AI Act**, which demand transparency, accountability, and robust governance frameworks.

The good news? You don't have to face these risks alone. Many organizations are already finding confidence in AI when it's deployed responsibly. According to a recent survey by Microsoft Security, 47% of current users of AI for security said they're very confident in its ability to make critical security decisions.[1] This shows that AI isn't just a tool for productivity and innovation—it can also be a trusted technology in protecting your business.

This guide is designed to help you get started. It's the first in a series of resources from Microsoft focused on security for AI applications, offering practical insights and strategies to help address today's most pressing risks. Future guides will dive deeper into areas like data security and compliance, but this issue focuses on laying the foundation for securing the AI tools your teams are already exploring.

Whether it's ensuring proper oversight of shadow AI, mitigating emerging threats like prompt injection attacks, or navigating evolving regulations, this guide provides the clarity and focus you need to take the first steps. With a phased approach grounded in Zero Trust principles, you'll gain the confidence to embrace AI securely and responsibly.

AI is a game changer—but only if you can secure it. Let's get started.

[1] Microsoft internal research, February 2025

By the numbers

Top 3 risks for securing AI applications

AI is unlocking incredible opportunities for organizations, but it's also introducing new risks that can't be ignored. Here are the top three big challenges every leader needs to tackle:

Data leakage and oversharing

of leaders fear sensitive information slipping through the cracks. Without proper oversight, employee use of unapproved tools **(shadow AI)** can expose sensitive information and increase the risk of breaches.

Emerging threats and vulnerabilities

of organizations worry about bad actors manipulating AI systems. New attack methods, such as prompt injection, **exploit vulnerabilities** in AI systems that are only now coming to light. As AI capabilities grow, so do the risks.

Compliance challenges

of leaders admit they're unsure how to navigate changing AI regulations. **Staying compliant** isn't just a box to check—it's critical to protecting innovation and avoiding costly setbacks.

[2] "First Annual Generative AI Study: Business Rewards vs. Security Risks," page 6, iSMG, 2024

[3] "Gartner Peer Community Poll: If your org's using any virtual assistants with AI capabilities, are you concerned about indirect prompt injection attacks?" Gartner

[4] "First Annual Generative AI Study: Business Rewards vs. Security Risks," Forrester, page 3, November 2024

The fundamentals

Setting the right security foundation for AI

AI relies on vast amounts of data to deliver results, but this reliance introduces unique risks. To help protect your organization, it's important to address three key areas: data leakage, emerging threats, and compliance challenges. Let's take a closer look.

Data leakage and oversharing

AI systems thrive on data, but too much access—or the wrong kind—can create vulnerabilities.

Shadow AI

You probably have experienced some early adopters of AI in your organization. They may believe that it is harmless if they "try it out" to make them more efficient—even if these tools aren't always approved by IT or security teams. For instance, a marketing team might use an unvetted chatbot to generate ideas, unintentionally exposing sensitive customer data by connecting to internal data sources without proper oversight.

To address this risk, you should implement centralized policies requiring employees to use only vetted tools. Deploy monitoring systems to detect unauthorized usage and ensure compliance with security standards. At the same time, provide secure, enterprise-grade alternatives so your teams can work efficiently without compromising sensitive information.

Over-permissioned data

When AI systems operate using the same permissions as the user, over-permissioning can expose sensitive data to misuse. A marketing analyst using an AI tool for customer insights, for example, might also have unnecessary access to financial records, which could inadvertently be included in the AI's analysis or outputs.

To reduce this risk, you should enforce role-based access controls (RBAC) to ensure employees access only the data they need for their roles. Establish clear policies around accessing data outside one's role, and monitor AI usage just as you would monitor search activity to detect and prevent inappropriate access.

Data lifecycle management

Poor retention policies and improper disposal of data can increase exposure and compliance risks, especially when AI tools access outdated or unnecessary information. For example, retaining customer purchase history beyond its regulatory retention period could lead to sensitive data being inadvertently accessed or processed by an AI tool, creating potential compliance violations.

You can minimize these risks by automating retention policies with lifecycle management tools, monitoring expiration timelines and enforcing secure deletion protocols to ensure compliance and protect sensitive information.

Emerging threats and vulnerabilities

As AI becomes more integrated into operations, new types of attacks are emerging.

Prompt injection attacks

Malicious inputs can manipulate AI systems into performing unintended actions or revealing sensitive information. Attackers, for instance, might embed hidden instructions in a document that instructs an AI system to send confidential information to an external server.

To help prevent such attacks, you should validate and sanitize all user-provided data before processing it. Limit the AI model's access to sensitive data and require users to verify their identity before interacting with high-risk systems. These measures will help ensure that only authorized individuals can interact with the AI securely.

AI errors: When AI makes mistakes

AI systems can make mistakes—just like humans do. These errors include **(1) hallucinations**—adding unsupported information; **(2) omissions**—leaving out critical details; **(3) bias**—skewed by race, gender, or other factors; **(4) garbage in, garbage out (GIGO)**—poor-quality or malicious input leading to flawed results; **(5) skewed focus**—emphasizing the wrong priorities; and **(6) overreliance**—blindly trusting AI outputs without question.

For example, an AI chatbot for customer self-service might confidently suggest an unsupported refund policy (hallucination), fail to mention key return exceptions (omission), favor certain customer demographics unfairly (bias), misinterpret poorly labeled training data to recommend irrelevant solutions (GIGO), focus too heavily on upselling products rather than resolving customer complaints (skewed focus), or be trusted without human review to resolve escalations (overreliance).

To help ensure AI applications are secure, organizations should implement robust monitoring and validation mechanisms to catch errors before they cause harm. Pre-built commercial AI tools often include safeguards like bias detection, input sanitization, and user access controls to reduce risks. Pairing these tools with strong oversight practices helps ensure that AI systems operate responsibly and securely while supporting business goals.

Compliance challenges

Unclear regulations and ethical concerns make it harder for organizations to adopt AI responsibly.

Policy development

Adapting AI systems to comply with evolving digital resilience regulations, such as the Digital Operational Resilience Act (DORA), is essential for ensuring operational continuity. For example, deploying AI-driven risk assessment tools without sufficient governance frameworks could result in noncompliance with DORA, exposing organizations to regulatory penalties.

You should establish governance frameworks aligned with DORA requirements, covering areas like resilience testing and ongoing monitoring. These frameworks should address AI-specific risks such as operational resilience, cybersecurity threats, and continuity planning.

Governance and documentation

Clear governance and thorough documentation are critical for deploying and implementing AI applications responsibly. Without proper records, organizations may struggle to track how AI applications are being used, increasing the risk of noncompliance with regulations like the EU AI Act.

To address this, you should maintain detailed documentation of how AI applications are deployed and monitored. This includes tracking data usage, model validation, system performance, and ongoing updates. Standardized processes and regular reviews will help ensure alignment with regulatory standards and support responsible AI implementation.

Clear governance and thorough documentation are critical for deploying and implementing AI applications responsibly

Automation

Assessing whether AI systems comply with regulations can be complex and error prone without the right tools. A financial institution deploying an AI-powered credit scoring system, for instance, may struggle to ensure compliance with General Data Protection Regulation (GDPR) requirements, risking noncompliance and regulatory penalties.

You can address this challenge by using AI-driven compliance platforms to continuously monitor AI applications for adherence to standards like GDPR or the Health Insurance Portability and Accountability Act (HIPAA). These tools can help evaluate risks unique to AI, such as data drift, unexplainable outputs, or unauthorized data access, ensuring ongoing regulatory alignment.

Making AI transparent and fair

Meeting ethical standards and privacy regulations requires transparency, fairness, and clarity in how AI applications operate. Under GDPR, for example, organizations must provide clear explanations for decisions made by AI that affect individuals, such as loan approvals or job screenings.

To meet these requirements, you should regularly review and document how AI applications make decisions. Conduct audits to ensure they operate fairly, respect individual rights, and comply with privacy regulations.

Navigating changing AI regulations

Rapidly evolving AI regulations can create uncertainty, making it challenging for organizations to stay compliant. Misclassifying the risk level of an AI application can lead to inadequate safeguards or wasted resources. For instance, incorrectly labeling an AI diagnostic tool as low risk could expose patients to harm and the organization to regulatory violations.

To avoid these issues, you should develop clear processes to assess and classify AI applications based on their risk levels. Staying informed about emerging regulations, such as the EU AI Act, will help ensure your practices remain aligned with current standards.

Spotlight on the EU AI Act

We're highlighting the **EU AI Act** because it sets a global benchmark for AI regulation, establishing strict standards for the safe, transparent, and accountable use of AI across the European Union. It provides clear guidelines for identifying and managing risks associated with AI technologies, requiring organizations to:

- Implement strong governance frameworks
- Maintain detailed documentation for data handling, model training, validation, and monitoring
- Ensure transparency, fairness, and explainability in AI decision-making processes

By aligning with the EU AI Act, you can proactively manage compliance, foster trust with customers, and stay ahead in a rapidly evolving regulatory landscape.

The horizon

Securing agentic AI

Agentic AI represents the next frontier of innovation. These systems go beyond traditional AI by acting independently, making decisions in real time, and collaborating with other systems to achieve complex goals. From optimizing energy grids to coordinating fleets of autonomous vehicles, agentic AI has the potential to transform industries and solve problems once thought unsolvable. But with this potential comes new risks.

What is agentic AI?

Agentic AI systems stand out because of their autonomy, adaptability, and ability to collaborate:

Autonomy

They operate independently based on set goals or learned behaviors, such as smart thermostats optimizing energy use.

Real-time decision-making

They process inputs instantly to act in fast-changing environments, like trading platforms executing millisecond trades.

Collaboration

They work with other agents to achieve shared objectives, such as warehouse robots coordinating tasks to manage inventory.

Unique security risks of agentic AI

While agentic AI offers transformative potential, its autonomy introduces distinct security challenges. These challenges often stem from the system's independence, adaptability, and reliance on vast amounts of data. Below are some of the most pressing risks:

Hallucinations and unintended outputs

Agentic AI systems can produce outputs that are inaccurate, outdated, or misaligned with user intent, leading to unintended results.

For instance, a self-service AI assistant might incorrectly suggest a product is in stock when it isn't or provide outdated shipping timelines due to stale data. These errors can confuse customers, damage trust, and create additional work for human agents who must address the fallout.

Organizations can reduce these risks by regularly updating training data, monitoring outputs, and reviewing AI-generated responses to catch errors early. Escalation paths for complex cases ensure alignment with customer needs, while human oversight helps deliver consistent and trustworthy experiences.

Overreliance on AI decisions

Blind trust in agentic AI systems can create significant vulnerabilities when users assume their outputs are always accurate or secure. For instance, a financial analyst might rely on an AI assistant to generate insights but inadvertently expose sensitive company data by acting on flawed recommendations without proper review.

New attack vectors

The autonomy and decision-making capabilities of agentic AI create new opportunities for attackers to exploit vulnerabilities:

Operational risks: Attackers can manipulate agentic AI systems to perform harmful actions, such as crafting personalized phishing emails or executing unauthorized tasks. For example, adversaries have used agentic processes to profile employees and compromise systems via indirect prompt injection.

Systemic risks: When multiple agents collaborate, a single compromised agent can cause cascading failures across interconnected systems. For instance, a hacked warehouse robot could influence others in its fleet, disrupting operations across the supply chain.

Accountability and liability

Agentic AI systems often make decisions without direct human oversight, which raises complex questions about accountability and liability. For instance, if an AI diagnostic tool in healthcare makes an error, it could lead to incorrect treatments or missed diagnoses—leaving organizations to grapple with who bears the legal and ethical responsibility for the mistake.

The playbook

Get started with a phased approach

With new AI innovations emerging in the market, such as agents, it is recommended to establish a strong foundation based on Zero Trust principles for securing AI applications from the outset. Built on the principle of "never trust, always verify," this approach ensures that every interaction is authenticated, authorized, and continuously monitored. However, achieving Zero Trust requires time; therefore, a phased approach allows for steady progress while building confidence in your organization's ability to securely integrate AI.

The AI adoption framework starts with fundamentals around AI Strategy and Planning. Once you have established what your strategy and goals are, then you need to plan out the scenarios for each area of your organization. This is where security and business teams need to work together, and you can get started with three key phases that set you in a strong position: Govern AI, Secure AI, and Manage AI. By focusing on these areas, organizations can establish a strong foundation for responsible AI use while addressing critical risks.

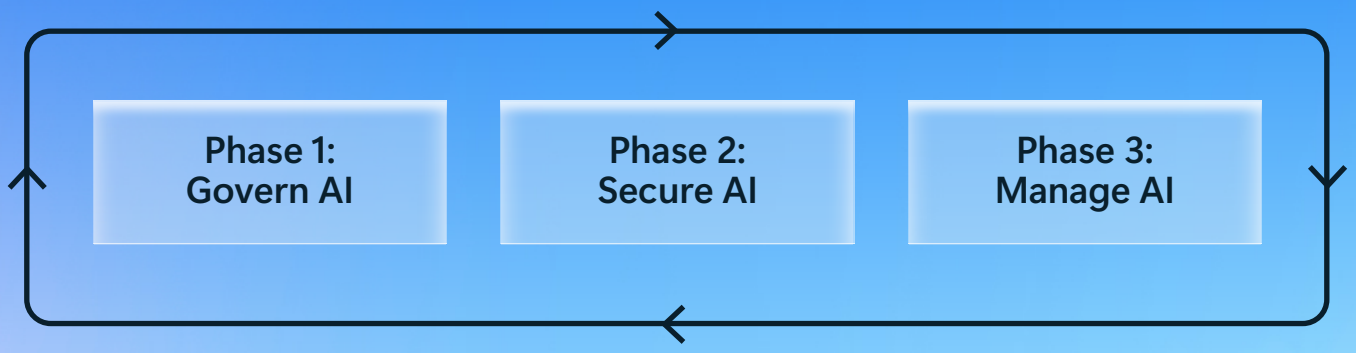


Figure 1: AI Adoption Guidance - AI adoption - Cloud Adoption Framework | Microsoft Learn

Phase 1: Govern AI

Start by creating governance frameworks to control how AI is used across your organization. This includes defining policies for responsible AI use, assessing risks tied to AI workloads, and enforcing guidelines to align with ethical standards, regulatory requirements, and business objectives. Automate policy enforcement where possible across AI deployments to help reduce the risk of human error. Regularly assess where automation can improve policy adherence.

> Learn more about Governing AI

Phase 2: Secure AI

Once governance is in place, prioritize securing AI systems to protect sensitive data, maintain model integrity, and ensure availability. Implement robust security controls, monitor for emerging threats, and conduct regular risk assessments to safeguard your AI environment.

> Learn more about Securing AI

Phase 3: Manage AI

Finally, focus on managing AI workloads effectively. This involves maintaining AI models, monitoring performance, and ensuring that systems remain reliable over time. Standardized practices and regular evaluations are essential to prevent issues like data drift or performance degradation.

> Learn more about Managing AI

This phased approach not only simplifies implementation but also helps ensure that your organization starts off with the right foundation needed to adopt AI responsibly and securely, one step at a time.

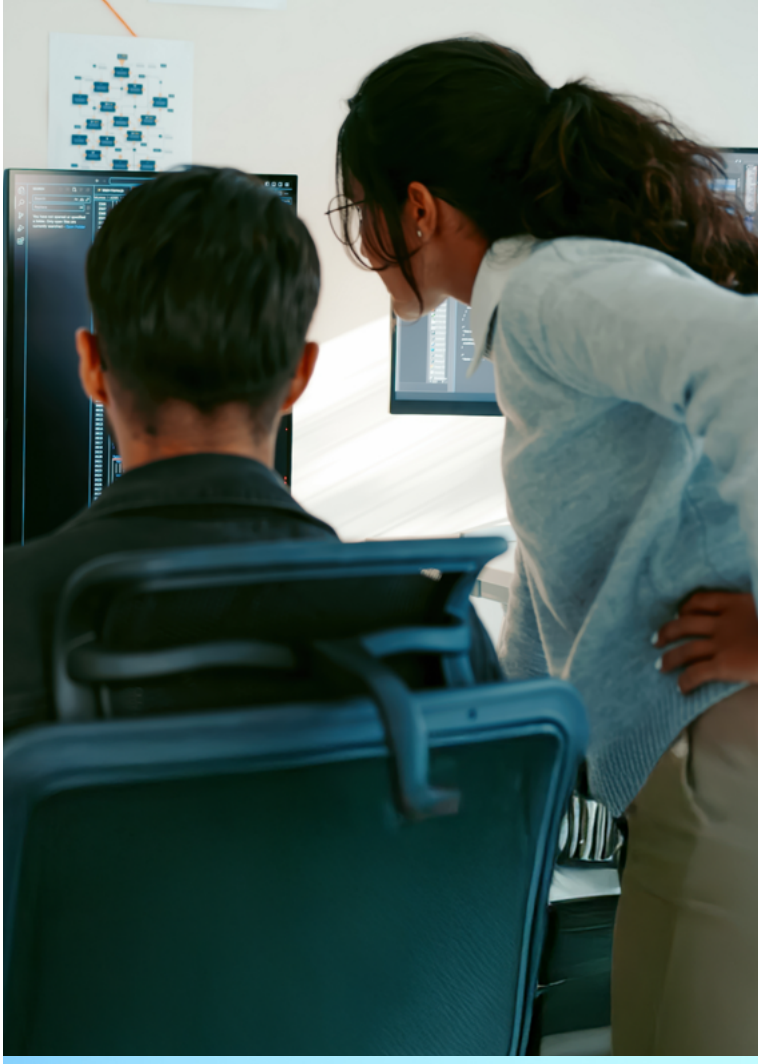
Building a foundation for success

AI is revolutionizing the way we work, unlocking unprecedented opportunities for growth and efficiency. Yet, it also introduces new risks that demand immediate and strategic action. By adopting Zero Trust as your foundation and following the AI adoption lifecycle—Govern AI, Secure AI, and Manage AI—your organization can secure AI systems while driving meaningful innovation.

Whether you're just beginning to explore AI or deploying advanced autonomous systems, fostering collaboration and embedding responsible practices will position you to navigate the evolving AI landscape with confidence.

To succeed, prioritize people. Train employees to recognize the risks associated with AI systems and provide clear guidance on secure usage practices and help them enable the AI applications that are IT and security approved. Encourage collaboration by breaking down silos between IT, security, and business teams to ensure a unified approach to AI security. At the same time, promote transparency by openly communicating your organization's AI security initiatives to build trust with stakeholders, strengthen relationships, and demonstrate leadership in this rapidly evolving space.

With the right strategy—grounded in Zero Trust principles and starting with a phased approach—you can help mitigate risks, unlock innovation, and build resilience to meet tomorrow's challenges head-on.



For more actionable cybersecurity insights, visit Security Insider.