Microsoft Security

AI security and Zero Trust

Agile security for agile business

In this whitepaper

Executive summary

Artificial intelligence (AI) transforms both the assets that require safeguarding and the mechanisms by which cybersecurity functions, adapts, and performs.

AI's potential to drive business forward is immense, but without proper management, overlooking security can undermine business initiatives, marketing campaigns, and regulatory compliance. By integrating security early and embracing Zero Trust principles, organizations can take advantage of AI while mitigating risks, much like brakes on a car enable people to safely travel faster.

In particular Generative AI (GenAI) is disrupting business operations and forcing everyone to evolve, including security teams. Most security leaders have encountered teams eager to implement new AI tools, sometimes without approval. Consider the sales manager who wants faster pricing information or the developer eager to use new tools without security blocking innovation. Everyone wants to maximize productivity, but many often overlook that security is essential to prevent damaging data leaks and other security risks.

Security teams must rethink their approach to protecting data and assets with AI. Because AI focuses on data, it's crucial to prioritize data classification and security. Traditional network defenses like firewalls simply can't effectively protect data and AI applications. Instead, you must protect data and assets wherever they are in the cloud, AI services, mobile devices, or anywhere else. The best way to do this is by adopting a comprehensive security framework like Zero Trust.

Key takeaways

- AI introduces multiple strategic imperatives for security
- Securing AI requires Zero Trust's asset-centric and data-centric approach
- AI requires evolving security controls and defenses
- AI increases the value, risks, and security needs for data
- AI can be used to help accelerate Zero Trust

This whitepaper explores how Zero Trust helps you navigate the security challenges and opportunities presented by AI so you can build a resilient and innovative digital infrastructure. Note that these learnings generally apply to multiple types of AI capabilities (including machine learning) but this paper primarily focuses on GenAI because of its power, popularity, and direct interaction with end users.

Key takeaways for AI Security and Zero Trust

While the security implications of AI are still emerging space, several learnings have become crystal clear. This whitepaper describes those key learnings:

Zero Trust and AI have a symbiotic relationship where they depend on each other

- AI requires Zero Trust, using an asset-centric and data-centric approach to secure both AI applications and their underlying data, as opposed to relying on a traditional network perimeter-centric security model.
- AI accelerates Zero Trust security modernization by enhancing security automation, offering deep insights, providing on-demand expertise, speeding up human learning, and more.

Rapidly update security strategy

Organizations must quickly adapt their security strategy because of the rapid adoption of GenAI by both attackers and business teams.

AI increases focus on data

AI is fundamentally a data analysis and generation technology, so the quality and security of AI applications is heavily reliant on the quality, lineage, classification, and protection of the underlying data.

Securing AI is a shared responsibility with the organization's provider

AI technology, like cloud technology, is most often a partnership with the provider that requires each partner to work on different aspects of security. It is crucial to learn this shared responsibility model and plan your security investments around it to effectively mitigate AI security risks.

Security controls need to be adapted to AI

Most existing security controls are built for classic deterministic computing that generates the exact same output for the same request each time. Many AI technologies dynamically generate new outputs each time, which requires updating existing security controls and introducing new ones to be effective.

The guidance in this whitepaper is designed to help you navigate the continuous changes posed by AI, capitalize on the opportunities, and manage the security risks and challenges.

Chapter 1

Strategic imperatives for security strategy

Attackers and business units are already adopting and using AI right now.

Security teams must acknowledge this reality and urgently update security strategy to enable adoption of the skills, processes, and tools to manage these risks effectively. The top strategic imperatives for AI security are:

Protect AI applications and data

Attackers are already targeting AI applications to steal data and to establish beachheads for larger attacks. You should integrate security experts into the development of AI-enabled applications to protect them from the beginning (as it will be much more expensive and difficult to fix security later). Security leaders should also work with business and technology leaders to shape the AI strategy to favor SaaS and PaaS to avoid the organization taking on unnecessary risk from "build your own" AI. See the [AI shared responsibility model] for more information.

Provide guidance to users

Attackers are already using AI to increase the quality and volume of existing attack techniques like phishing emails, scam phone calls for business email compromise, and more. Review your use policy, user support processes, and user education to ensure users are aware of how convincing attacker communications can be, how to identify these threats, and how to escalate them to security teams.

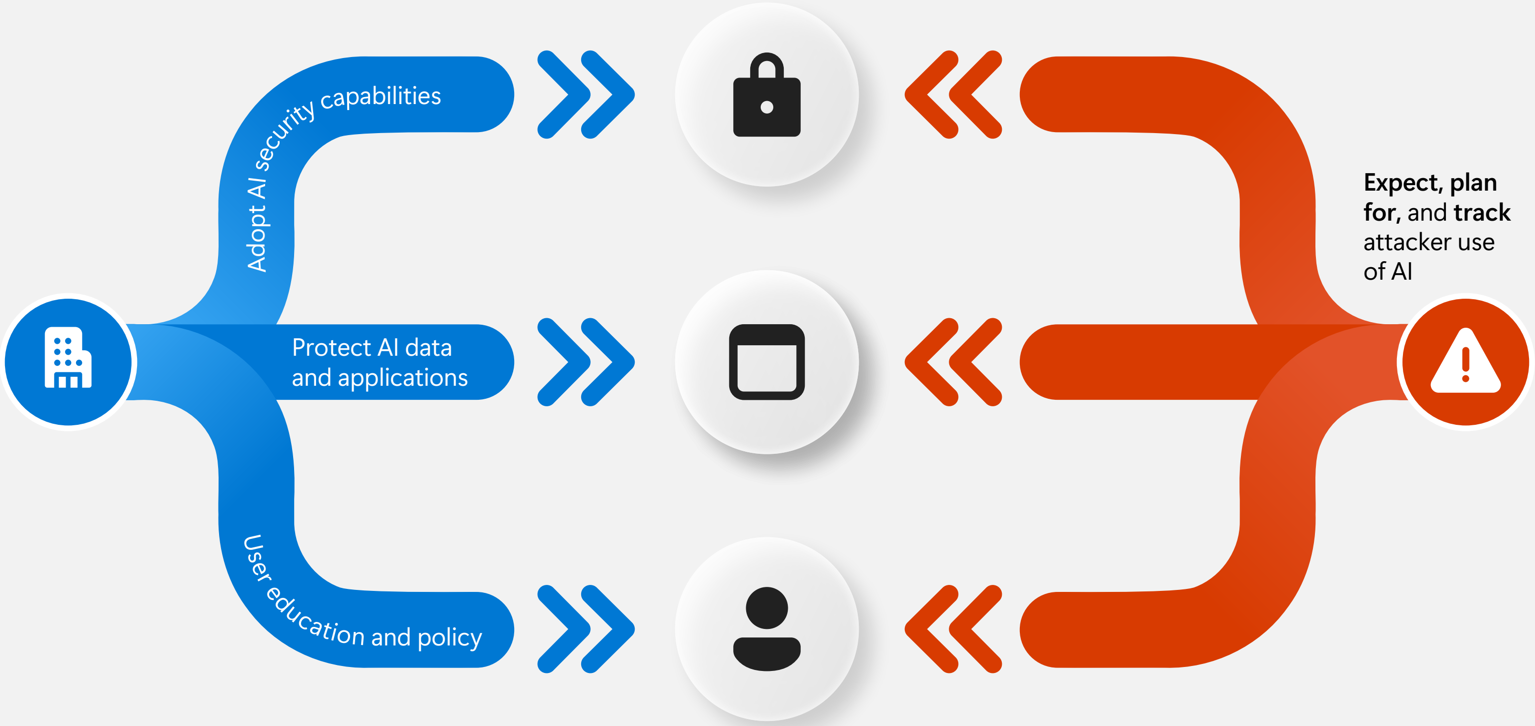
Adopt AI security capabilities

AI technology is no silver bullet, but it provides clear and compelling value in key scenarios like guiding analysts through the incident response process, summarizing the impact of an attack, building reports on incidents and investigations, and reverse engineering scripts. Security teams should evaluate AI security capabilities to see if it will increase their ability to keep up with attacks

Establish appropriate standards

Organizations should ensure they have written standards that can guide organizational decisions and show due diligence and due care to regulators and other 3rd parties. These standards typically cover security, privacy, and ethical topics depending on the organization's expected and authorized use of AI. For an example you can use Microsoft's Responsible AI Standard as a reference.

Managing multiple dimensions of AI security risk


Adopt AI security capabilities
Protect AI data and applications
User education and policy
Expect, plan for, and **track** attacker use of AI

Chapter 2

Zero Trust and AI: A symbiotic relationship for end-to-end security

AI is changing what security needs to protect and how it operates

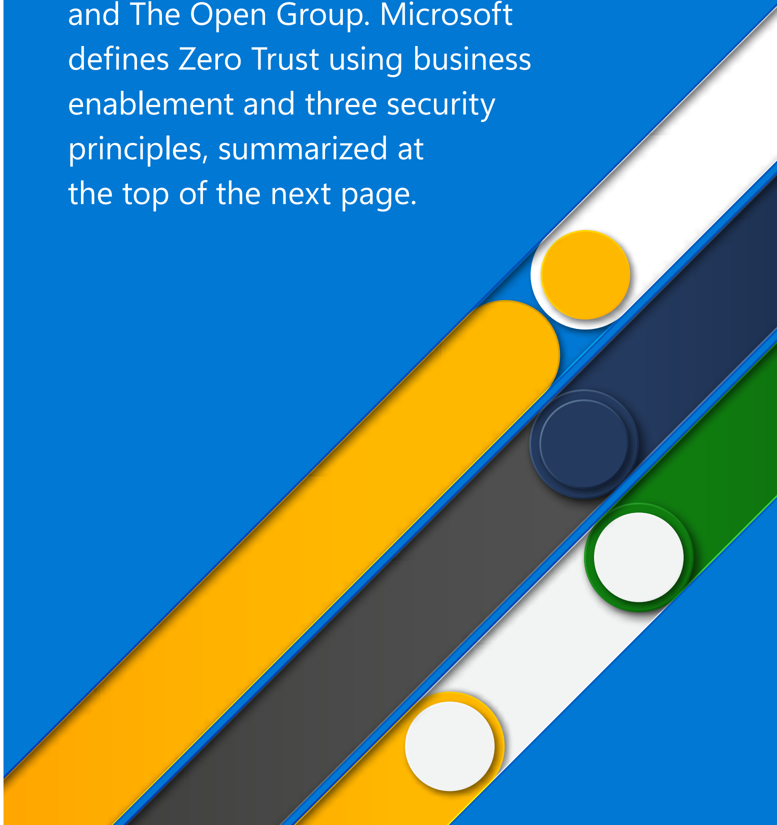
AI and Zero Trust work together in a powerful, symbiotic relationship. AI thrives within the secure environment that Zero Trust creates, while Zero Trust evolves to meet the challenges that AI introduces. Each enhances the other, helping to ensure that your security framework is both adaptive and resilient.

This Zero Trust transformation from a classic network-centric approach to an asset- and data-centric approach is recommended for effective security in today's age of cloud services, mobile devices, and now AI. General infrastructure like network configurations, firewalls, and access controls form the foundation, but these are not sufficient on their own. Just as cars require specific safety measures, digital assets need data classification, data encryption, adaptive access management, and other asset-centric security measures to stay safe.

What is Zero Trust?

Zero Trust is a modern security approach that aligns to business risk and real world attacks, continuously adapts, and pragmatically prioritizes investments.

Zero Trust has been defined by NIST and The Open Group. Microsoft defines Zero Trust using business enablement and three security principles, summarized at the top of the next page.



Verify explicitly

Protect assets against attacker control by explicitly validating that all trust and security decisions use all relevant available information and telemetry.

Use least privilege access

Limit access of a potentially compromised asset, typically with just-in-time and just-enough-access (JIT/JEA) and risk-based policies like adaptive access control.

Assume breach

Assume attackers can and will successfully attack anything (identity, network, device, app, infrastructure, etc.) and plan accordingly.

Business enablement: Align security to the organization's mission, priorities, risks, and processes.

Considering Zero Trust to help secure AI

Classic network-perimeter security approaches simply cannot protect data or AI applications. Firewalls, IDS/IPS, and Network DLP controls in a network perimeter focus on detecting and mitigating risk using static predictable patterns of network traffic. These classic approaches don't work effectively for AI applications because:

- **AI network traffic is often encrypted** for privacy and security reasons, which prevents these network-based controls from getting any visibility.
- **AI operates at data and application** abstraction layers, so differentiating between dangerous and safe actions/communications requires controls that understand applications, data, and users.
- **AI activity is dynamic,** so it doesn't match static patterns. These controls are designed to detect and block. Additionally, AI also allows rapid generation of new tooling that can copy existing functionality (e.g. a custom NMAP clone) which can evade static signature-based defenses.

Additionally, one way to protect data effectively is with a data-centric approach that stays with data wherever it goes. Network controls are limited to the organization's security perimeter and the assets in it so they cannot protect data on mobile devices, cloud services, USB drives, and other locations. For these reasons, a Zero Trust asset-centric and data-centric approach is an effective way to protect AI and related data assets.

AI accelerates the adoption of Zero Trust

AI can play a pivotal role in accelerating the implementation and operationalization of a Zero Trust strategy. Generative AI enhances this by acting as an on-demand resource for learning and automating key tasks, such as data discovery and classification. This not only speeds up workflows but also ensures that data is managed securely and consistently across the organization.

Note that many recommended data security and governance practices have been known before the usage of the 'Zero Trust' name, but they still fit the Zero Trust philosophy perfectly. These have also often been deferred or deprioritized at many organizations, so they are often 'new' approaches to the organization.

Generating automation (scripts, programs, etc.) also accelerates Zero Trust by allowing security teams to avoid repetitive work and focus their efforts on strategic activities instead. This also reduces repetitive manual tasks, which are prone to human error and can cause risk. Note that these should be created in a secure by design manner to avoid introducing additional security risks.

Furthermore, Generative AI can enhance both business process understanding and security risk identification. For example, it can help organizations discover sensitive data and assess whether unauthorized individuals have access, while also detecting patterns that indicate unusual data movement within or outside the organization. These patterns could point to insider risks, innocent errors, or even external data exfiltration attempts. By learning from these patterns, AI not only benefits productivity but also refines security design, helping business and security teams better understand processes and identify vulnerabilities.

Chapter 3

Data value, risks, and security needs are amplified by AI

AI elevates focus on classifying and protecting data

AI massively amplifies the priority of data security for an organization. Organizations often recognize the importance of data security but have frequently had to defer or deprioritize it in favor of more urgent priorities like modernizing identity and access security for the cloud, maturing security operations, adapting infrastructure and development security practices to cloud and DevOps, or other key initiatives.

The advent of AI means organizations must prioritize data security to help tackle the important (and challenging) work of classifying and protecting their data. This increase in prioritization is primarily driven by two factors:

- AI increases the value of data (to businesses and attackers)
- AI amplifies existing data security and governance challenges

AI increases the value of data (to business and attackers)

GenAI's ability to generate insights from data has transformed it into an even more valuable asset, as AI models increasingly become a core driver of business profitability. These models require high-quality, original data sources for training, further elevating the importance of proprietary data. As a result, enterprise data is not only crucial for business success but also a lucrative target for cyber attackers.

GenAI's success depends heavily on the quality, lineage, classification, and protection of the data it processes. With the increasing saturation of low-quality data on the open internet, public data sources have become less reliable for training robust AI models. This makes high-quality enterprise data an increasingly valuable resource, not only for building better AI but also as a prime target for cyber attackers. These attackers seek to exploit enterprise data for financial gain, either by using it to train their own models or by selling it to other malicious actors.

Furthermore, the integration of AI with disparate data sets can lead to blurred lines around data ownership and stewardship, complicating the already complex issues of privacy and intellectual property. Accidental disclosures, whether through training models or in retrieval-augmented generation (RAG) applications, pose significant risks to organizations. As AI continues to evolve, safeguarding enterprise data becomes critical, not only for producing reliable AI outputs but also for ensuring that this valuable asset isn't compromised or weaponized by external threats.

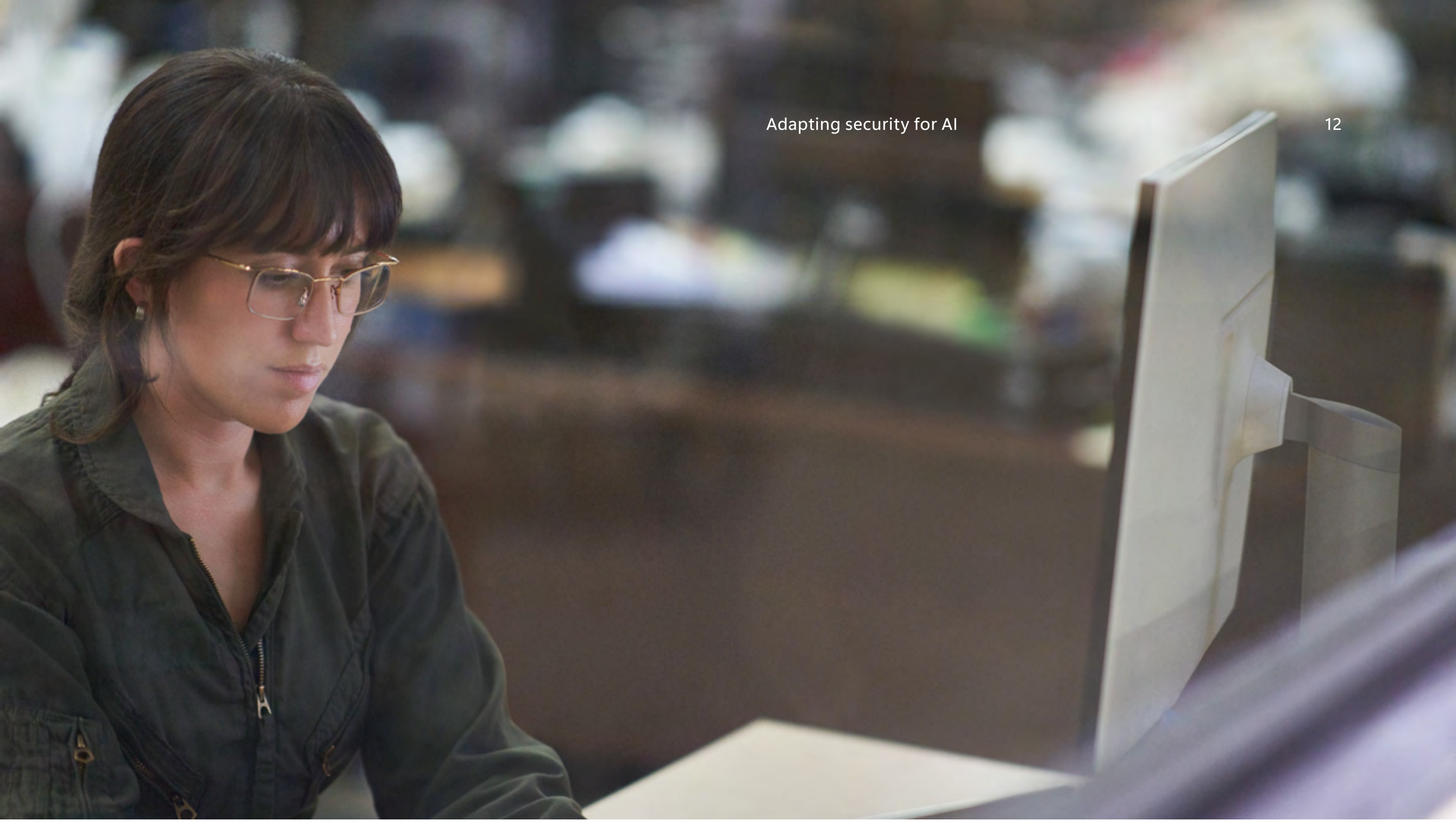
AI amplifies data security and governance challenges

One of the strengths of GenAI is the ability to discover data and make it easily accessible. Because many organizations have not established or consistently applied a formal data classification strategy, introducing AI initiatives can make sensitive information easily discoverable to unauthorized users (which they weren't previously aware of). While users may have access to these documents already, they often don't know how to find them or have the time to search through them.

For example, internal users may ask an AI application about the salaries and compensation for other employees and executives in the organizations or external attackers may ask AI what secret projects the organization is working on. The AI may provide answers to those unauthorized users if the documents describing employee compensation and sensitive projects have not been classified correctly, or the AI application does not recognize or enforce access rules based on these classifications. For the record, Microsoft Copilot respects your identity model and permissions, inherits your sensitivity labels, applies your retention policies, supports audit of interactions, and follows your administrative settings.

Organizations must recognize these challenges and update their data governance and security strategies, starting with clear policies and procedures for data classification. These must be supported with technical controls on the data itself as well as the applications that use the data (including AI applications).

As AI reshapes business roles and elevates the importance of data, these changes create both opportunities and risks. However, it's not all bad news as AI can also help discover and mitigate data risks in addition to the value it creates for business.



Chapter 4

Adapting security for AI

A shared responsibility model for AI security

Securing AI systems is a partnership where security responsibility is shared between organizations and their AI providers, similar to cloud technology. It is critical for all stakeholders to learn this shared responsibility model and plan their security investments, strategies, and controls based on this model. This collaborative approach helps keep AI systems secure and resilient against evolving threats.

The three layers of an AI system are as follows:

1. **AI platform**
   Microsoft offers a range of AI solutions running on Azure, including those powering their own Copilot solutions.

2. **AI application**
   Software developed by the organization to ensure productive and secure use of generative AI solutions.

3. **AI usage**
   How generative AI is used within an organization, including data consumption and generation.

The table below summarizes the shared responsibilities between organizations and AI providers in securing AI applications across these layers. Depending on the type of AI deployment—Infrastructure as a Service (IaaS), Platform as a Service (PaaS), or Software as a Service (SaaS)—the division of responsibilities changes:

IaaS

The organization builds their AI models on a cloud platform like Azure, where Microsoft provides the infrastructure. The customer manages the security of their models, data, and applications.

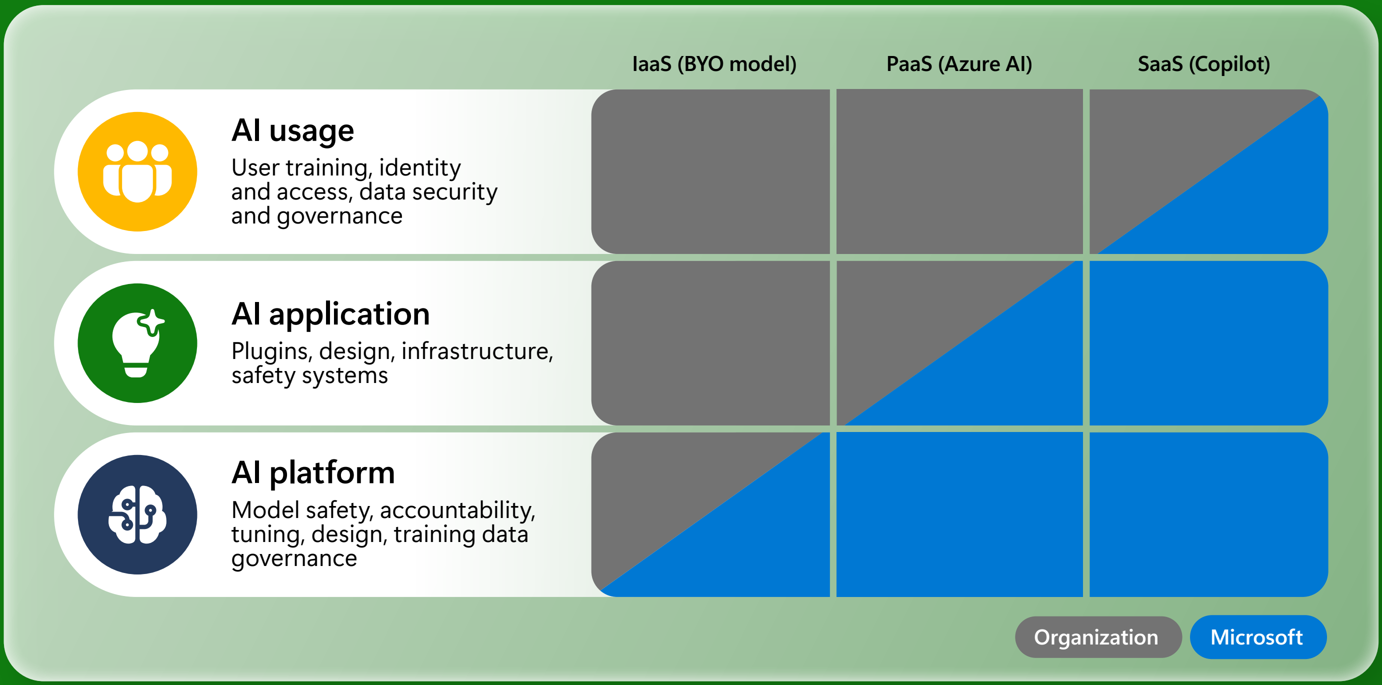
PaaS

The customer develops applications on top of Azure AI offerings, with Microsoft providing many embedded controls. The customer is responsible for securing the custom application and its usage.

SaaS

Managed services, and/or Microsoft's Copilot can help deliver the necessary functionality without the customer needing to develop or manage software. The customer still manages how the service is used and secures any data provided or generated.

AI security shared responsibility model



	IaaS (BYO model)	PaaS (Azure AI)	SaaS (Copilot)
AI usage User training, identity and access, data security and governance	⬛	⬛	⬛/🟦
AI application Plugins, design, infrastructure, safety systems	⬛	⬛/🟦	🟦
AI platform Model safety, accountability, tuning, design, training data governance	⬛/🟦	🟦	🟦

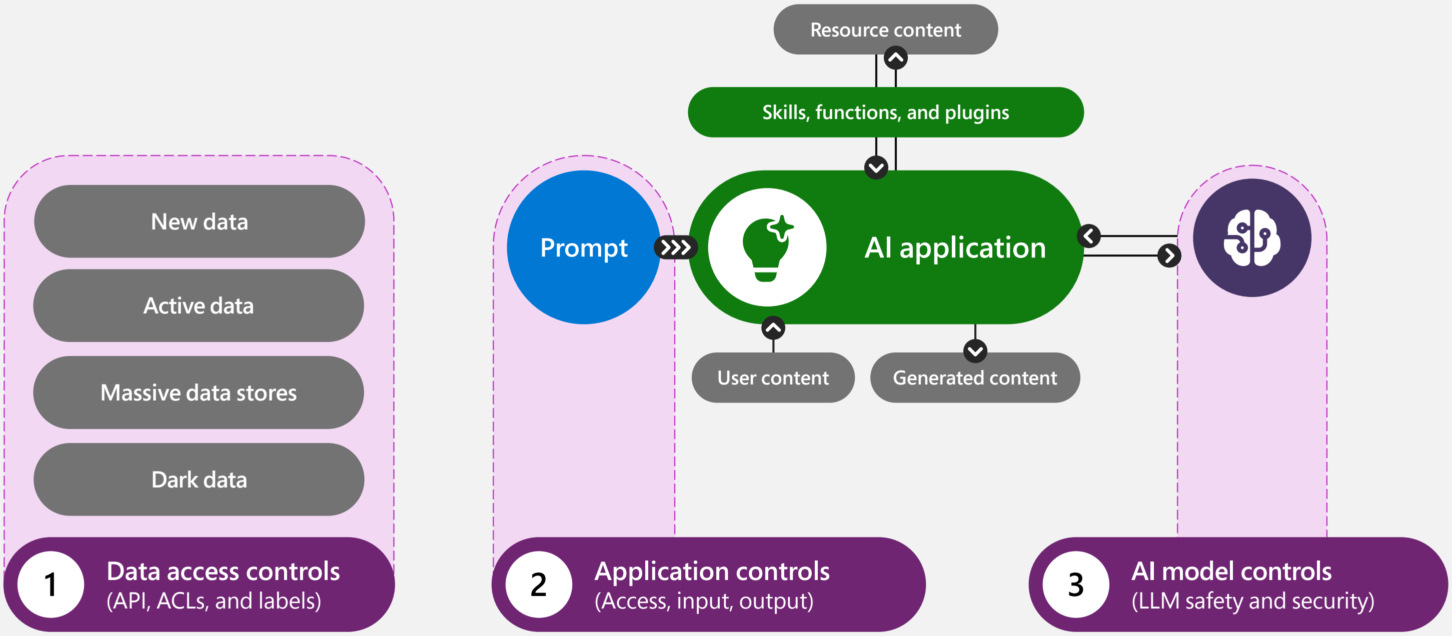
⬛ Organization 🟦 Microsoft

For a more detailed version of this model, see AI shared responsibility model.

Operationalizing a shared responsibility model

To operationalize this shared responsibility that safeguard AI at every level, organizations can focus on three key areas, as shown in the diagram below. These three pillars form the foundation to implement these controls and help create resilient AI systems.

1. **Data access controls:** Safeguard data with APIs, ACLs, and labeling.
2. **Application controls:** Manage how applications interact with data and models.
3. **AI model controls:** Ensure AI models are secure to prevent unintended disclosures.



Don’t undo these boundaries
Ensuring that access to data is strictly regulated through mechanisms like APIs, Access Control Lists, and data labeling. This helps maintain the integrity and confidentiality of data, preventing unauthorized access or misuse.

Don’t give unlimited access
Managing how applications interact with data and models, including the regulation of input, processing, and output. This prevents AI applications from becoming a weak link in the security chain, especially when dealing with sensitive or critical information.

It will give the secrets it knows
Safeguarding AI models, particularly large language models, to prevent them from inadvertently revealing sensitive information or being manipulated to produce harmful outputs. These controls are vital in maintaining the trustworthiness and security of AI systems.

AI-specific security measures

AI applications are fundamentally different than traditional applications. Traditional applications are deterministic, which means they generate the exact same output every time they get the same input. Today's security controls and security assumptions are built around that predictability.

AI based applications that use generative AI models are different because they are dynamic in nature—the model will generate a different output each time they are run with the exact same input. For example, asking an image generation model to "draw a picture of a kitten in a security guard uniform" repeatedly is unlikely to generate the exact same picture twice (though they will all be similar).

This dynamism offers new value for businesses but also introduces new types of security risks. This dynamism also means that current (deterministic) security controls designed will not be effective against AI applications. This requires the organization to rethink their data practices and security controls to ensure safe use of AI:

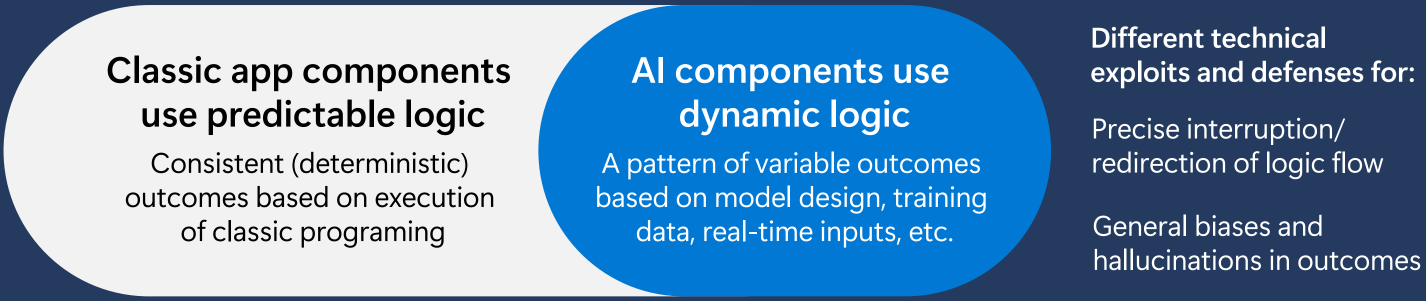
Attack simulation (red team/pen testing) has to operate differently, focusing on using human language to trick AI models in addition to exploiting deterministic code vulnerabilities.

Security and technology roles have to rely heavily on threat models to evaluate these new system designs until a knowledgebase of security controls are established for standard application patterns.

Business and AI application roles that are sponsoring and developing AI projects need to work with security teams to understand the inherent risks for AI and available mitigations.

Data owners need to work with security teams to ensure that sensitive data is classified and handled properly by AI (which may be excluding its use by AI).

The image below illustrates how AI applications are typically a combination of both predictable deterministic logic and dynamic AI logic:



Classic app components use predictable logic

Consistent (deterministic) outcomes based on execution of classic programing

AI components use dynamic logic

A pattern of variable outcomes based on model design, training data, real-time inputs, etc.

Different technical exploits and defenses for:

Precise interruption/ redirection of logic flow

General biases and hallucinations in outcomes

For more information on the types of threats to AI, see Microsoft AI Red Team

Chapter 5

Conclusion and futures

Without adopting a Zero Trust approach to security and integrating these learnings on AI, your organization could face increased risk of damaging cybersecurity attacks and diminished business returns from AI initiatives.

Here at Microsoft, we recommend organizations adopt a Zero Trust approach for security to give them the agility and asset-centric controls to manage risks to AI and data. Just as a transportation system needs more than road signs and barriers to keep people safe, an organization needs more than basic network security controls to protect business assets. Road signs and stripes are crucial, but you also need car-specific protection like insurance, seatbelts, airbags, and collision avoidance systems to make cars and drivers safe.

As we navigate the brave new world of continuous changes driven by AI, cloud, mobility, and more, we are confident that the Zero Trust principles can guide the way and help you navigate the opportunities and challenges yet to come.

Guidance and technical resources

The following resources expand on the principles, lessons learned, and requirements covered earlier to help accelerate your AI and Zero Trust readiness:

Security Adoption Framework (SAF)

Guidance on adopting Microsoft security solutions, helping organizations effectively implement and optimize security strategies.

Adoption Scenario Plan Phase Grid

Easily understand the security enhancements for each business scenario and the level of effort for the stages and objectives of the Plan phase.

Zero Trust adoption tracker

Downloadable PowerPoint deck to track your progress through the stages and objectives of the Plan phase.

Business scenario objectives and tasks

Downloadable Excel workbook to assign ownership and track your progress through the stages, objectives, and tasks of the Plan phase.

Additional Zero Trust documentation

See additional Zero Trust content based on a documentation set or your role in your organization.

Adversarial threat landscape for AI systems

Utilize MITRE's ATLAS™ to understand and defend against adversarial threats targeting AI systems.

Zero Trust for Microsoft Copilots

Apply Zero Trust protections to Microsoft Copilots.

Partner integration with Zero Trust

Apply Zero Trust protections to partner Microsoft cloud solutions.

Best practices for AI security risk management

Learn about best practices for managing AI security risks to protect your AI deployments.

Threat modeling AI/ML systems and dependencies

Explore comprehensive guidance on threat modeling AI/ML systems to identify and mitigate potential security risks.

NIST AI Risk Management Framework

Review the NIST AI Risk Management Framework for standards and guidelines to manage risks related to AI.

Strengthen your Zero Trust posture blog

Offering practical guidance for organizations to enhance their security posture with integrated, streamlined tools.