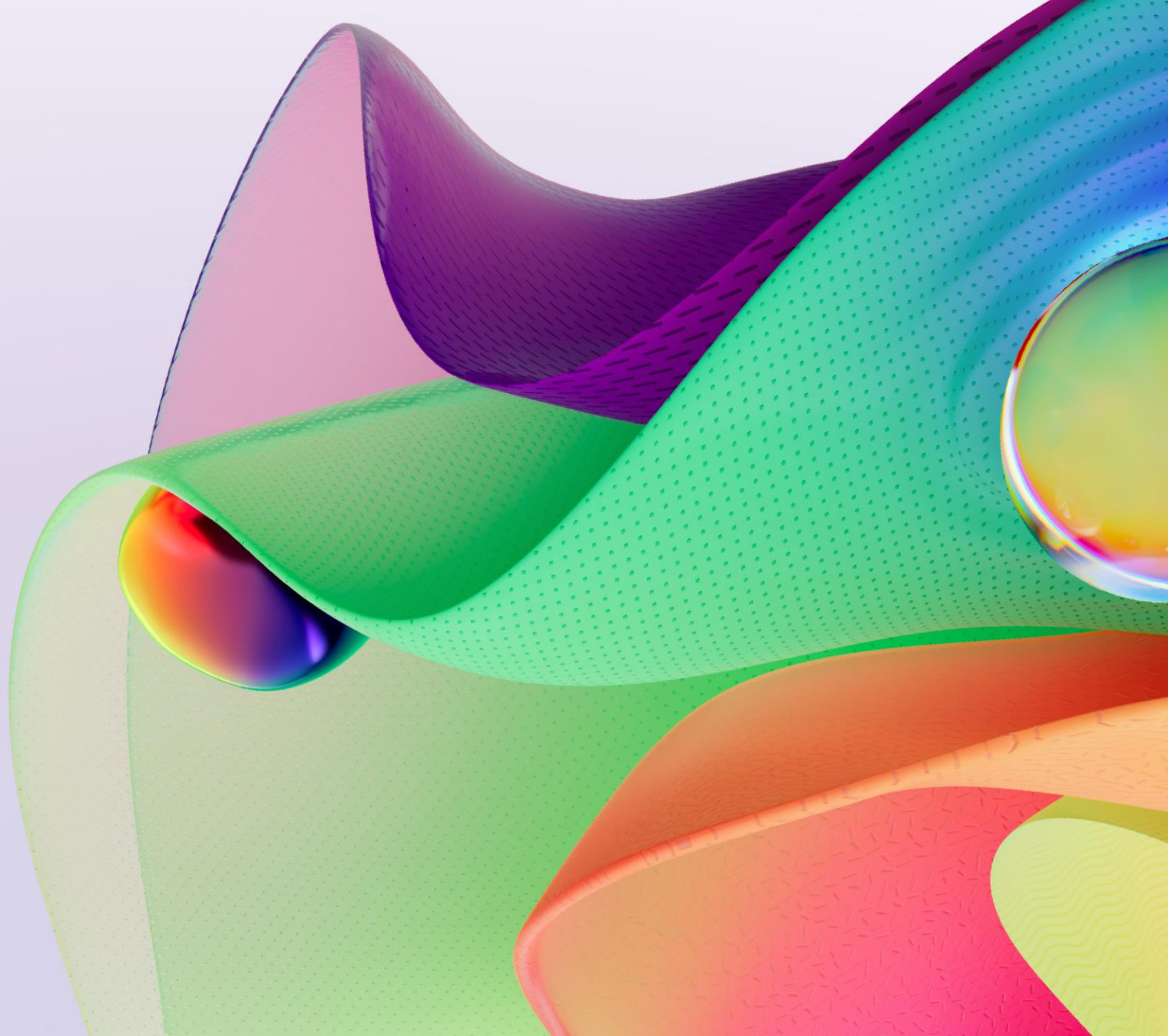




Microsoft AI Tour

In partnership with  **NVIDIA.**





Securing Generative AI Applications

Joylynn Kirui
Senior Cloud Security Advocate
Microsoft

X: @joylynn_kirui



Agenda



Introduction



Understanding AI architecture in a security context



The AI security threat landscape

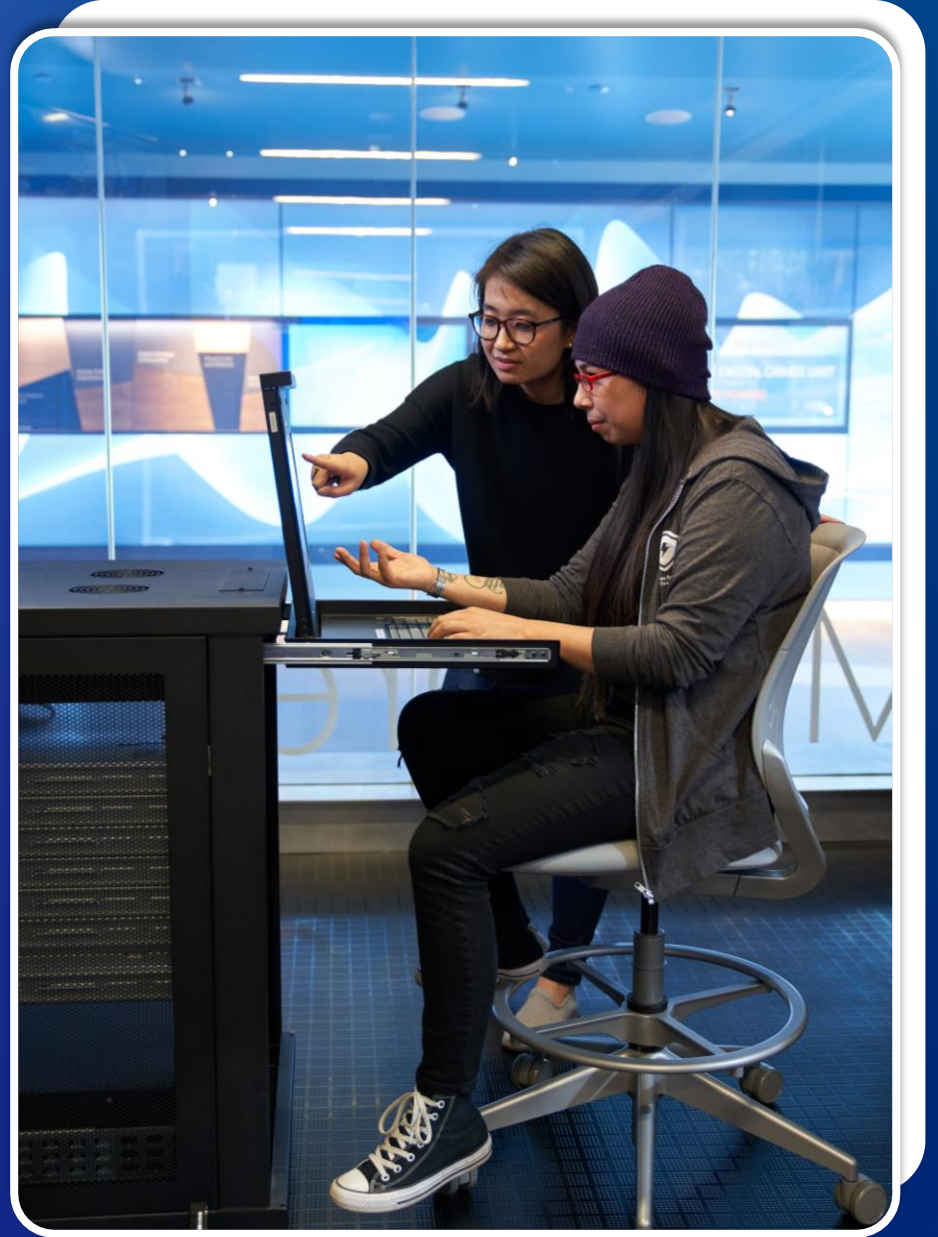


How Microsoft secures AI platforms

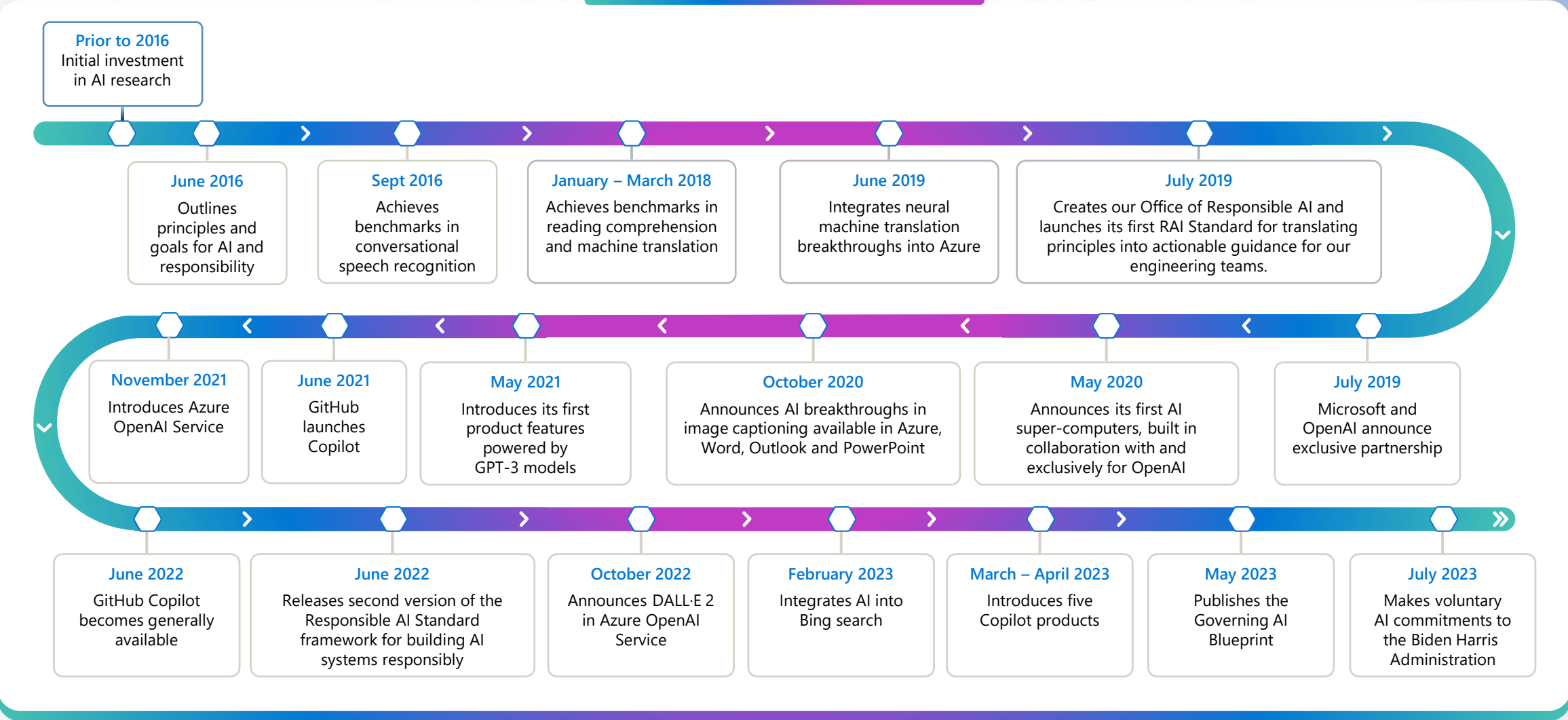


Security controls for developers building AI-enabled applications

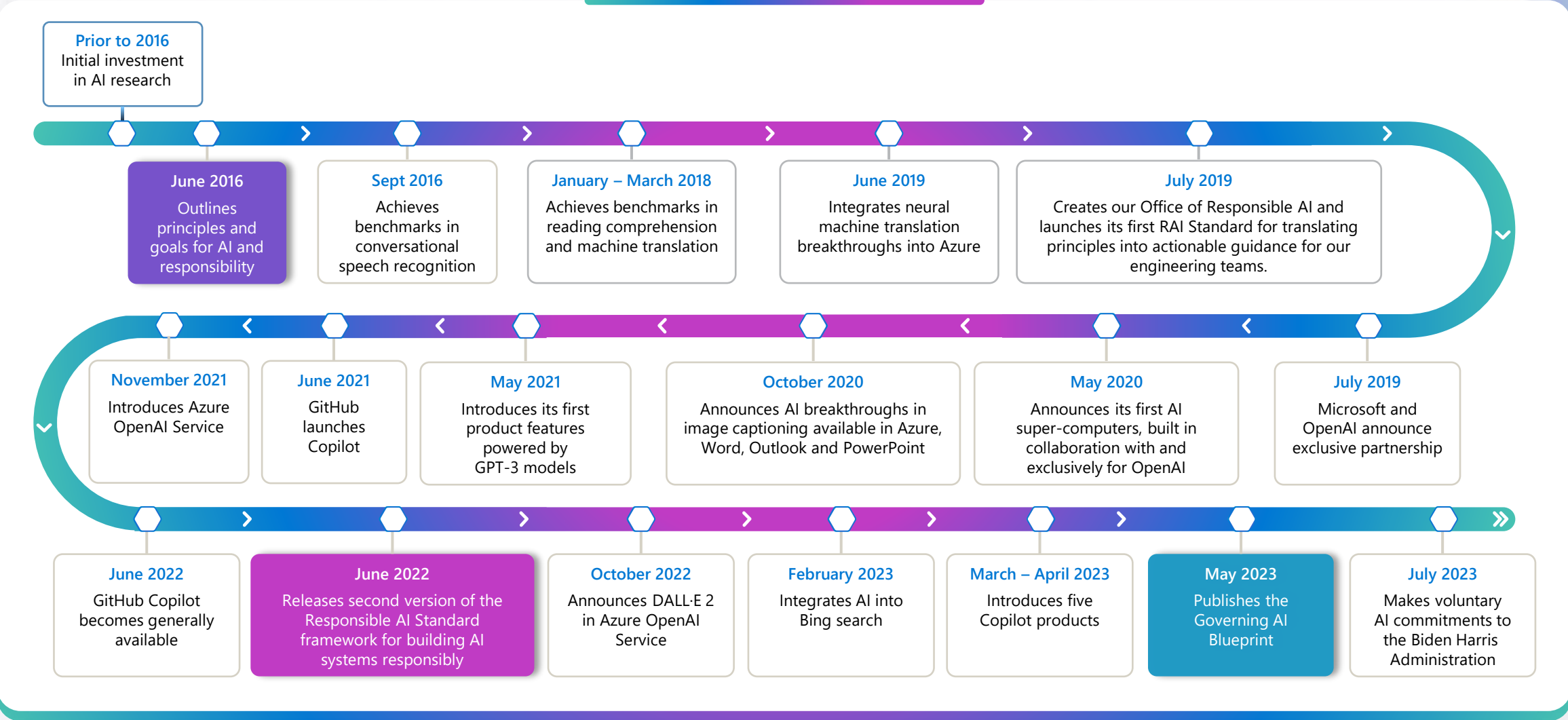
Introduction



Microsoft's Journey in AI



Microsoft's Journey in AI



Microsoft's Responsible AI Framework



Fairness

AI systems should treat all people fairly.

Reliability & Safety

AI systems should perform reliably and safely.

Privacy & Security

AI systems should be secure and respect privacy.

Inclusiveness

AI systems should empower everyone and engage people.

Transparency

AI systems should be understandable.

Accountability

People should be accountable for AI systems.

[Learn More](https://www.microsoft.com/en-us/ai/responsible-ai)

<https://www.microsoft.com/en-us/ai/responsible-ai>

Microsoft's Responsible AI Framework



Fairness

AI systems should treat all people fairly.



Reliability & Safety

AI systems should perform reliably and safely.



Privacy & Security

AI systems should be secure and respect privacy.



Inclusiveness

AI systems should empower everyone and engage people.



Transparency

AI systems should be understandable.



Accountability

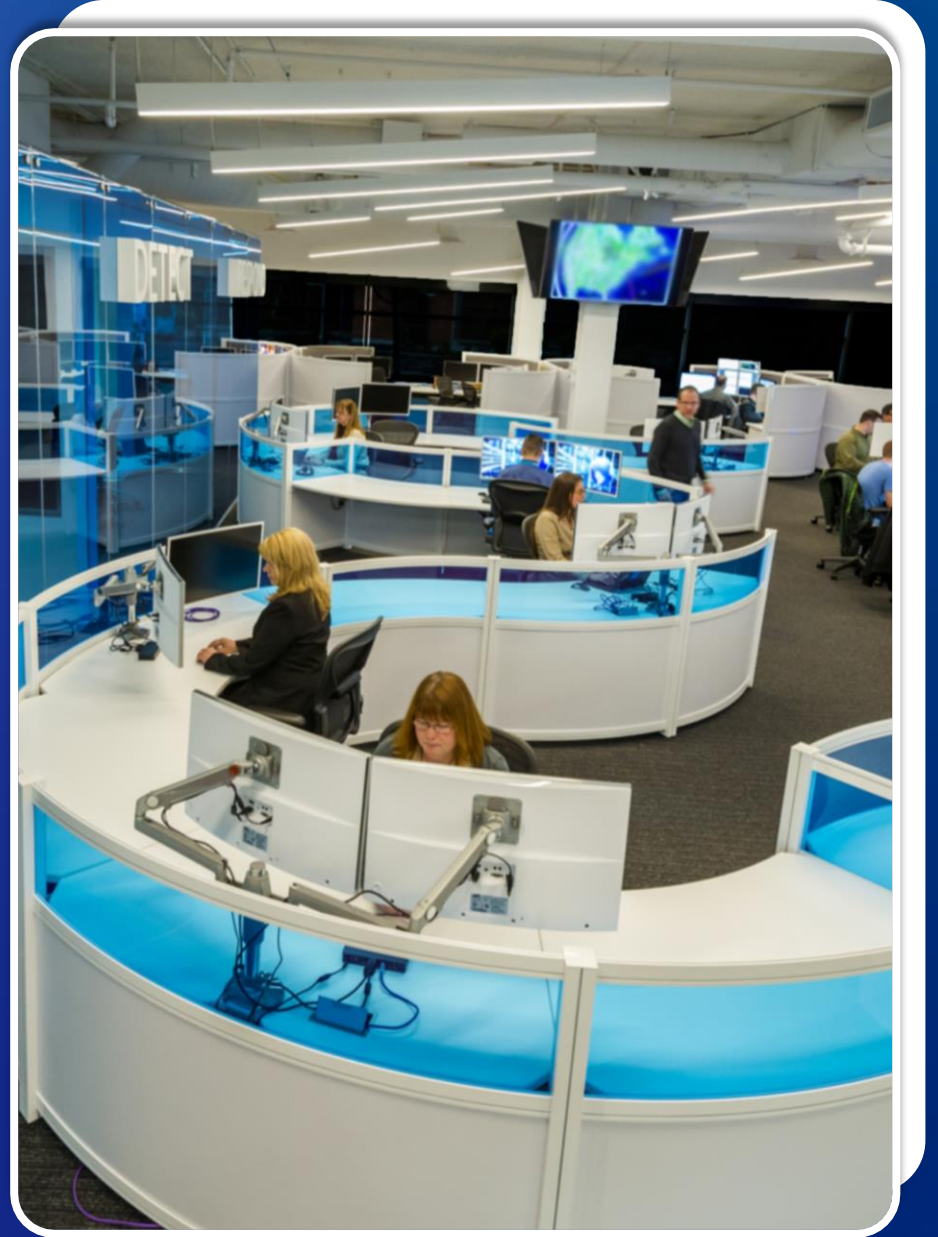
People should be accountable for AI systems.



[Learn More](#)

<https://www.microsoft.com/en-us/ai/responsible-ai>

Understanding AI architecture in a security context



AI Architecture Overview

AI Usage

This layer focuses on the user interaction with the AI interface, either standalone or built into existing application UI.

Current Examples:

- Public Access (Bing Chat(Copilot)/ChatGPT/Bard)
- Code Development (GitHub Copilot)
- Microsoft Copilots (Office, Viva, Windows)

AI Application

Development of an AI-integrated application through secured software development practices (SDL). Additional plugins are enabled to provide specific functions and controls.

Current Examples:

- Microsoft development of Copilots
- Customers developing their own products
- 3rd party solution offerings running on Azure

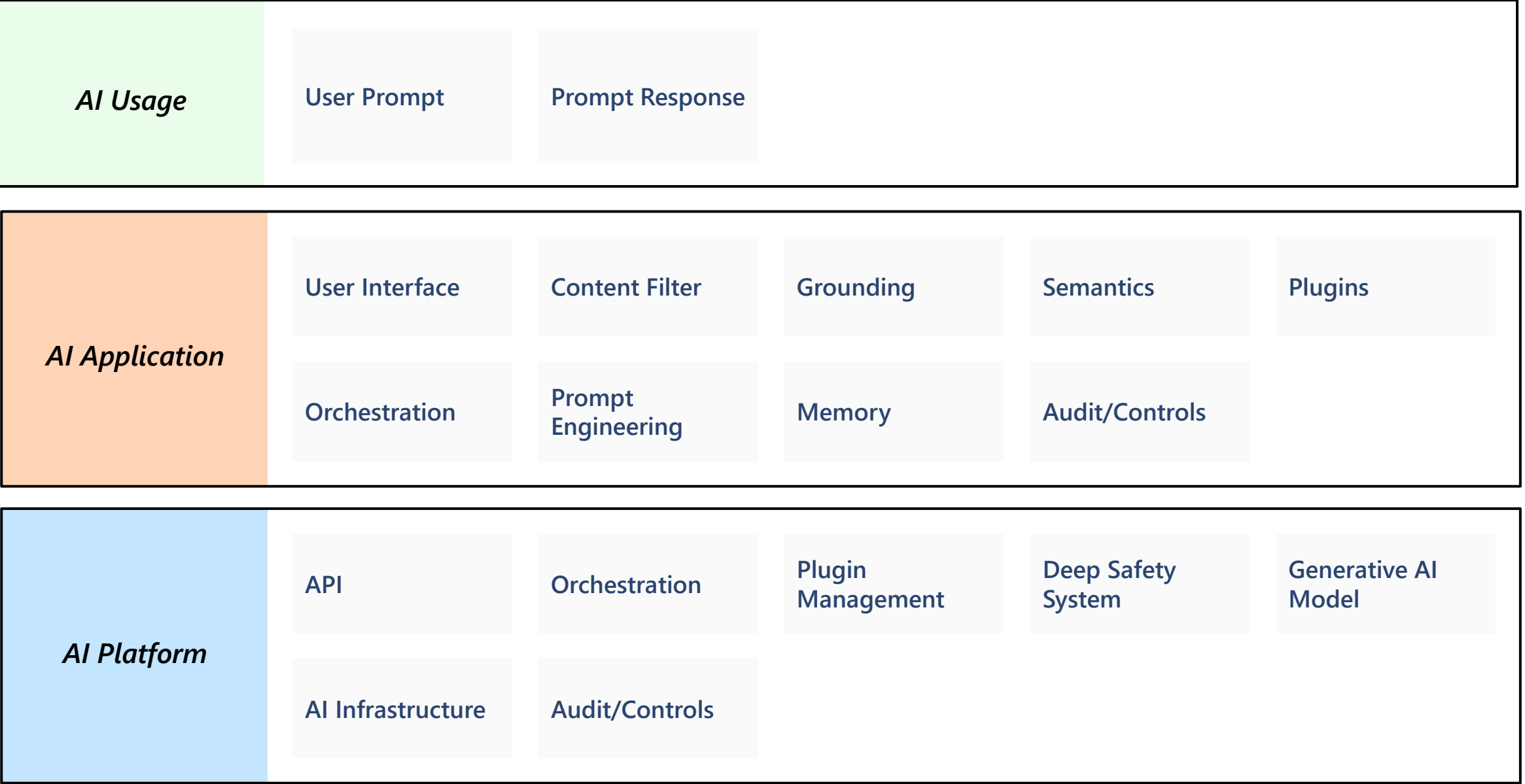
AI Platform

Access to traditional and generative AI models to build your own AI-integrated solutions, running in your compliance boundary on secure and reliable cloud infrastructure.

Current Examples:

- Azure ML Model Catalogue (Hugging Face, Llama 2)
- Azure OpenAI Service (GPT4, ChatGPT, DALL-E)

Detailed AI architecture breakdown



AI Shared Responsibility Model

Illustrates which responsibilities are typically performed by an organization and application developer and which are performed by their AI provider (such as Microsoft)

		IaaS (BYO Model)	PaaS (Azure AI)	SaaS (Copilot)
AI Usage	User training and accountability	Customer	Customer	Customer
	Usage policy, admin controls	Customer	Customer	Customer
	Identity, device, and access management	Customer	Customer	Shared
	Data governance	Customer	Customer	Shared
AI Application	AI plugins and data connections	Customer	Customer	Shared
	Application design and implementation	Customer	Customer	Microsoft
	Application infrastructure	Customer	Customer	Microsoft
	Application safety systems	Customer	Shared	Microsoft
AI Platform	Model safety and security systems	Customer	Shared	Microsoft
	Model accountability	Customer	Model Dependent	Microsoft
	Model tuning	Customer	Model Dependent	Microsoft
	Model design and implementation	Customer	Model Dependent	Microsoft
	Model Training Data Governance	Customer	Model Dependent	Microsoft
	AI Compute Infrastructure	Shared	Microsoft	Microsoft

Microsoft

Model Dependent

Shared

Customer

AI Pain Points



Integration

New technologies and design decisions introduce new risks and vulnerabilities.

User training needs to be adapted to the new capabilities of the AI solutions selected for use by the organization.



Data & Privacy

Sensitive data access and processing via AI systems creates new risks.

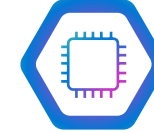
Transparency and control needs to be established and maintained throughout the lifecycle.



AI Supply Chain

Increased focus on potentially vulnerable or malicious code or 3rd party components.

Lack of compliance standards and rapidly developing best practices.



Trusted AI

Very similar to the early days of BYOD: Employees likely already using GenAI to achieve their tasks.

Leaders must establish a trusted pathway to GenAI integrated applications to protect the organization.



The Unknowns

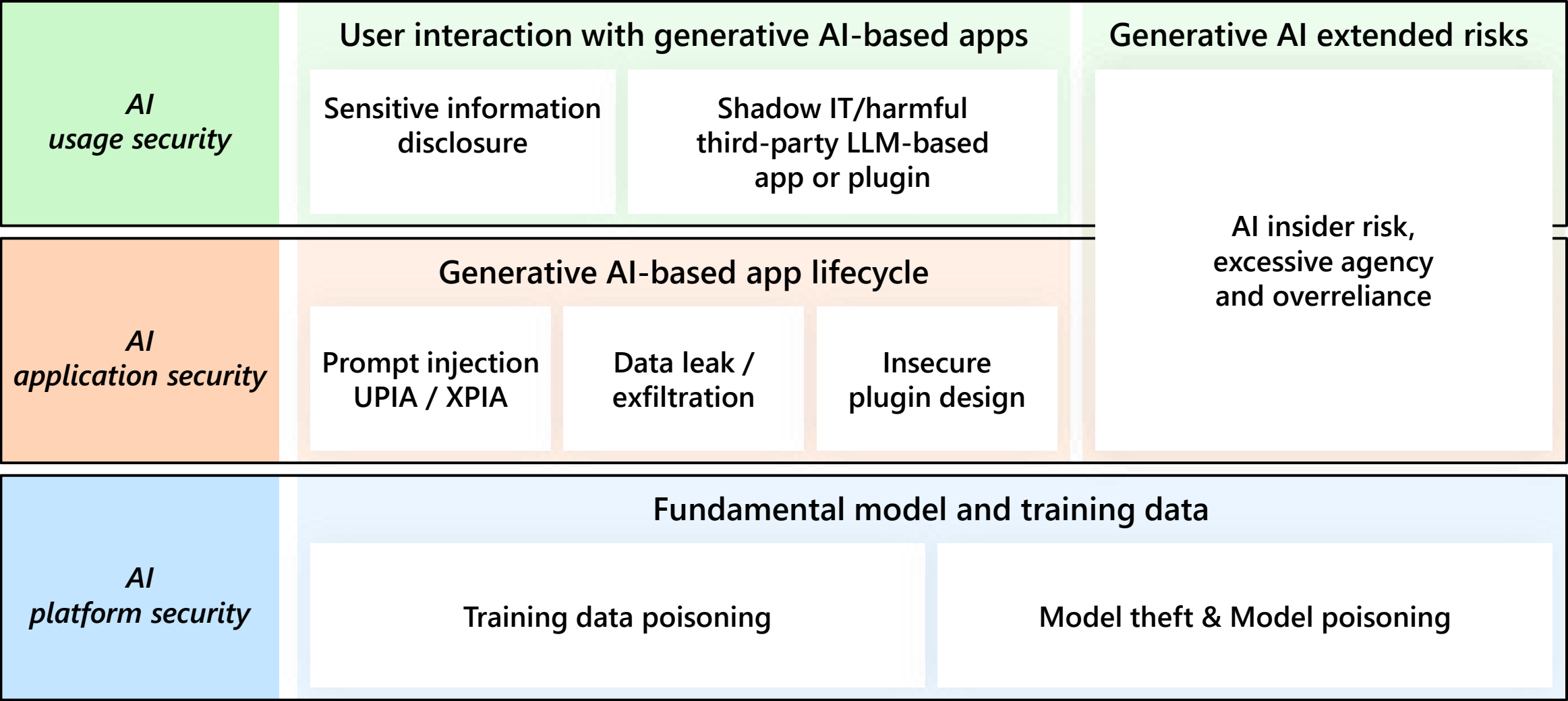
GenAI is new and brings unique challenges such as AI Hallucinations.

Exciting prospects of the potential, the ROI is not yet proven in real-world scenarios.

The AI security threat landscape



Generative AI threat map



The cybersecurity bell curve

Basic security hygiene still protects against 98% of attacks¹



Enable multifactor authentication

Make it harder for bad actors to utilize stolen or phished credentials by enabling multifactor authentication. Always authenticate and authorize based on all available data points, including user identity, location, device health, service or workload, data classification, and anomalies.

Apply least privilege access

Prevent attackers from spreading across the network by applying least privilege access principles, which limits user access with just-in-time and just-enough-access (JIT/JEA), risk-based adaptive policies, and data protection to help secure both data and productivity.

Keep up to date

Mitigate the risk of software vulnerabilities by ensuring your organization's devices, infrastructure, and applications are kept up to date and correctly configured. Endpoint management solutions allow policies to be pushed to machines for correct configuration and ensure systems are running the latest versions.

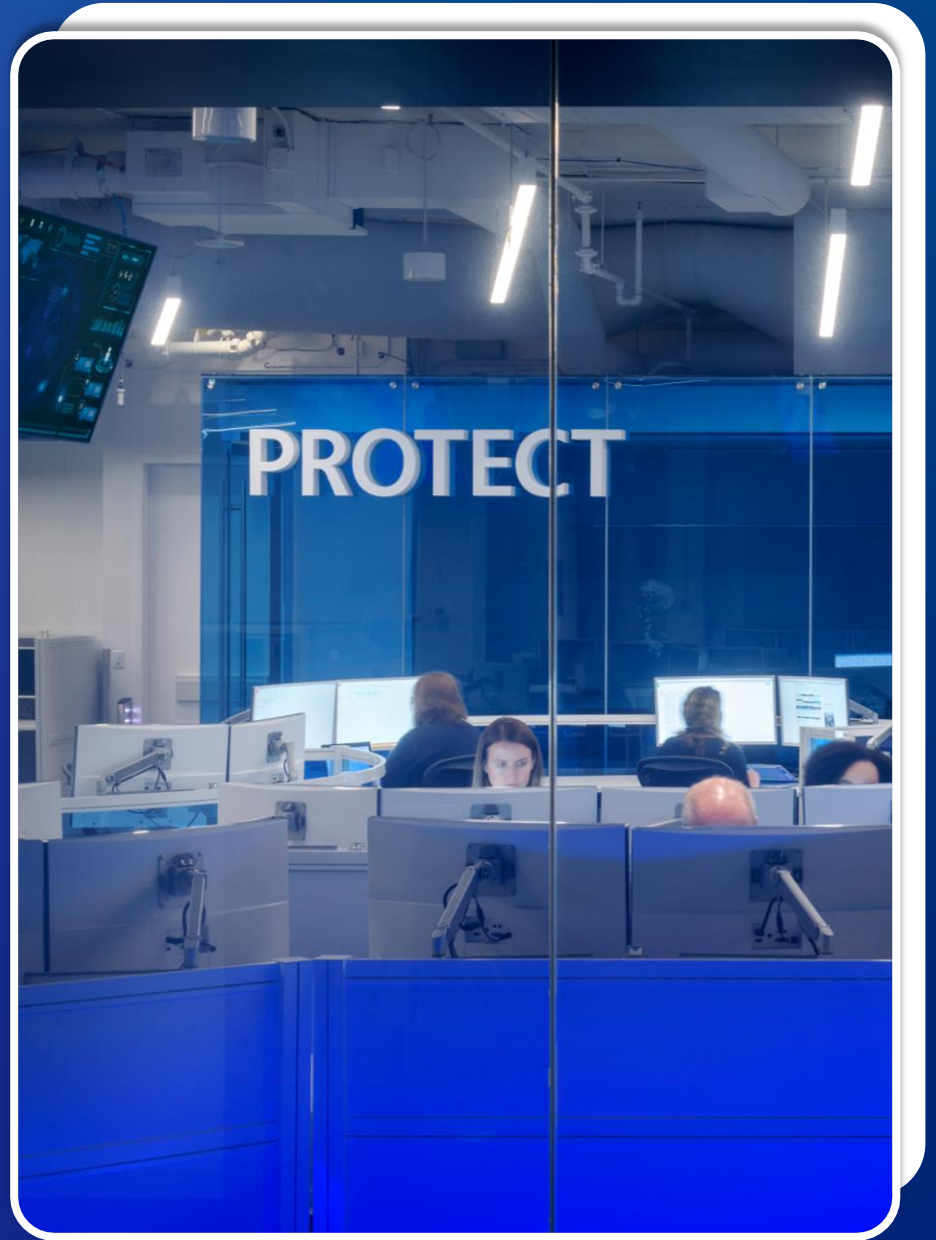
Utilize antimalware

Stop malware attacks from executing by installing and enabling antimalware solutions on endpoints and devices. Utilize cloud-connected antimalware services for the most current and accurate detection capabilities.

Protect data

Know where your sensitive data is stored and who has access. Implement information protection best practices such as applying sensitivity labels and data loss prevention policies. If a breach does occur, it's critical that security teams know where the most sensitive data is stored and accessed.

How Microsoft secures AI platforms



Building new principles for AI security

Ensure your data is *your* data



Microsoft will not use customers data to train the foundational AI models, without explicit consent. Data governance is a shared responsibility.

Customer:

Protect your data as a top priority. Ensure it remains private and controlled, end to end.

Be secure by design



The Microsoft AI stack is designed and built on decades of secure software practices, and a strong supply chain of partners, following mature SDL tools and processes.

Customer:

Ask for transparency in every AI system you connect to your data, for the whole AI supply chain.

Secure by intention and in practice

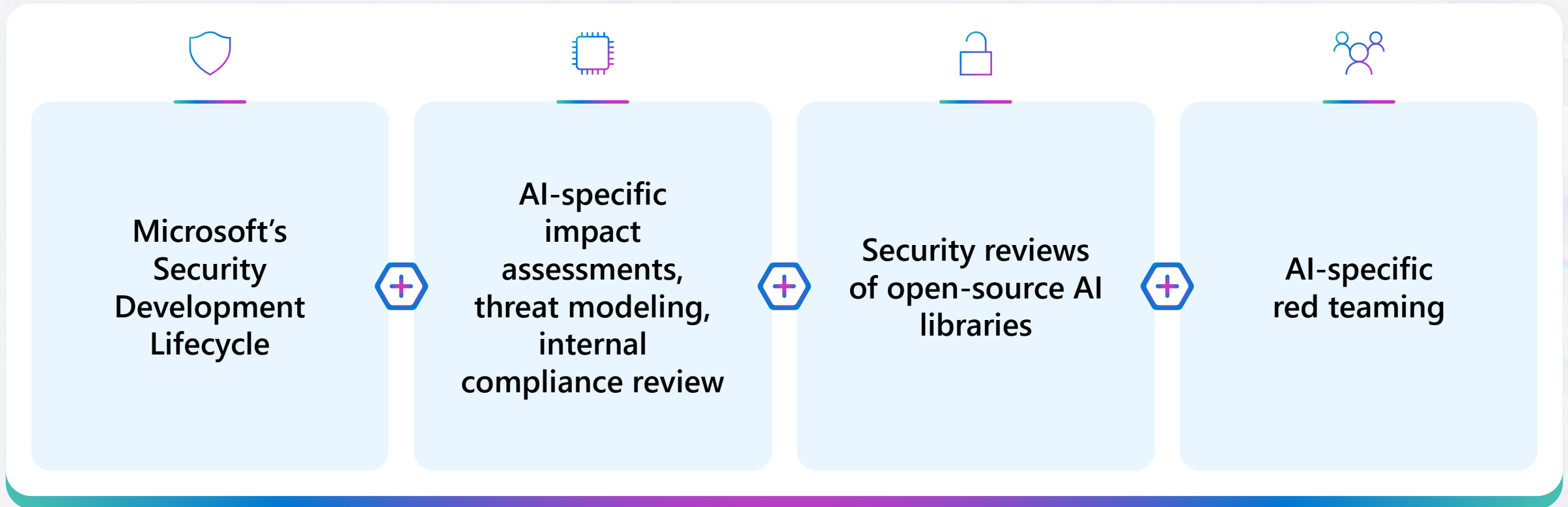


Microsoft grounds its efforts to advance AI in strong principles that govern the technology's implementation and use: Positioning people at the center of the equation.

Customer:

Strong Zero Trust, and Data Governance programs will matter more than ever.

How does Microsoft address risks from attackers?



How does Microsoft address risks from the use or misuse of AI?



Example: Overreliance

Using AI to justify a viewpoint or action

Assuming the AI must be fair or accurate

AI doing something that the user
can't meaningfully check.

User is simply too busy to check it carefully



Microsoft Approach

Ground on authoritative data sources

Provide greater transparency and explainability

Design UX interfaces to mitigate overreliance

How does Microsoft address risks from the use or misuse of AI?



Example: Hostile Misuse

Hostile misuse involves using AI system to intentionally to cause harm including circumventing safeguards.

Generating
malicious code

Asking instructions for
harmful purposes

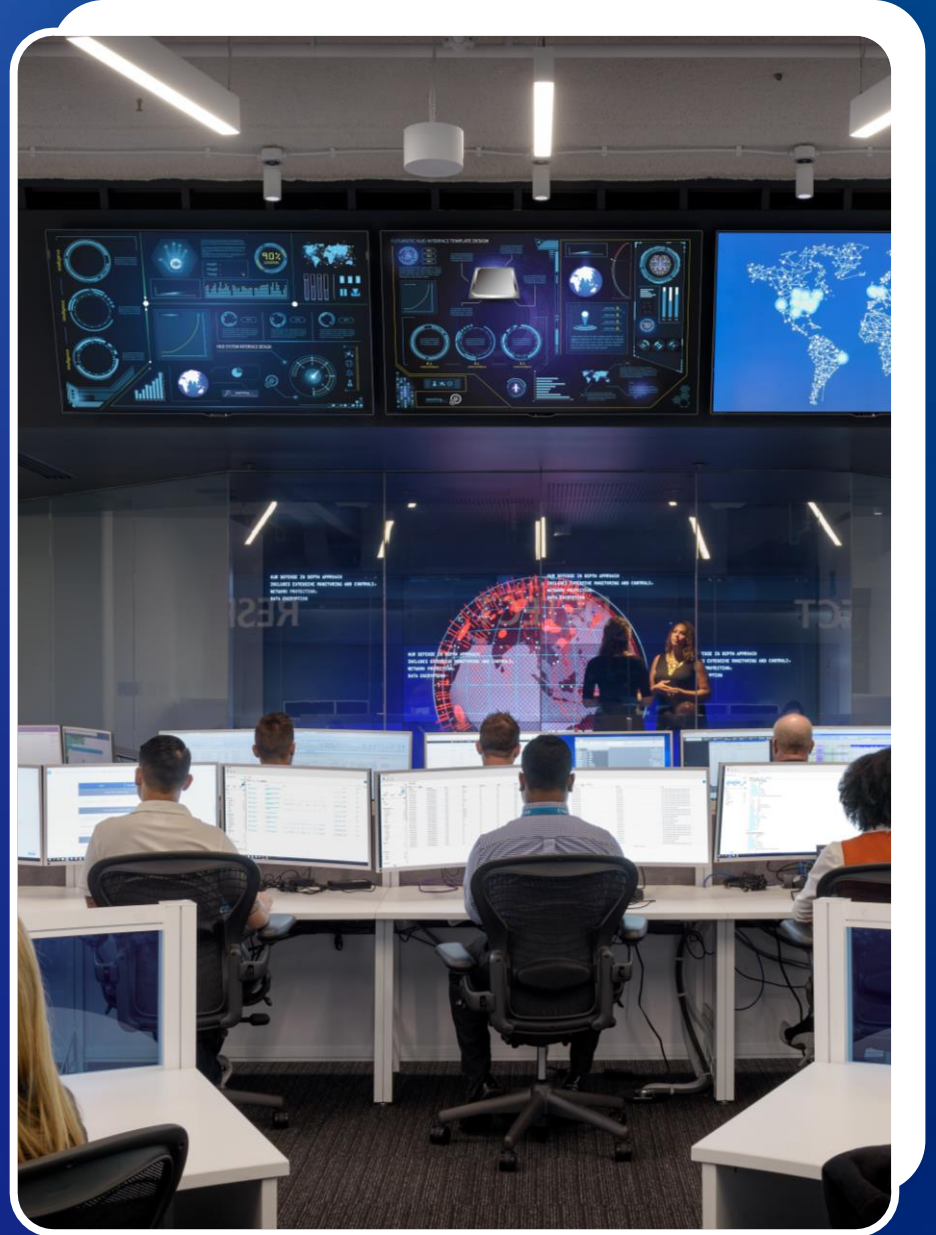


Microsoft Approach

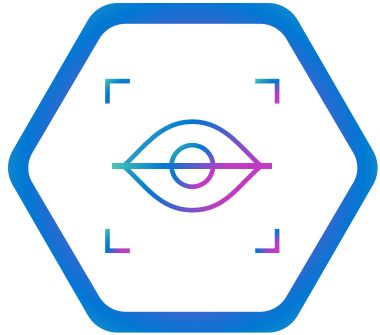
Microsoft has invested heavily in a defense in depth approach including a deep safety layer that provides security by default to disallow the AI to perform tasks that are harmful or dangerous to the user, intentionally deceptive, or likely to adversely affect the public interest

Acceptable Use Policy governs AI usage

Microsoft's AI Red Teaming Approach



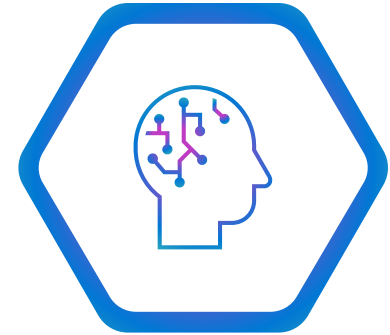
Security Community view of Red Teaming



Double Blind

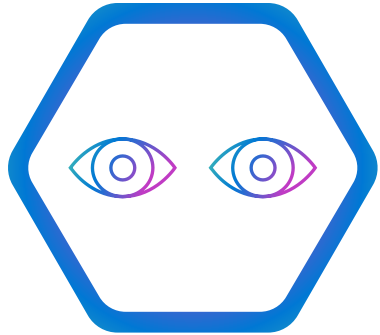


Emulate real
world adversaries

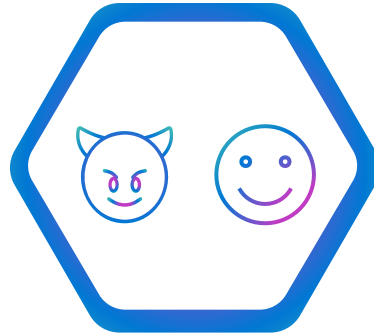


Mature toolkit
and processes

RAI Community view of Red Teaming



Generally
single blind

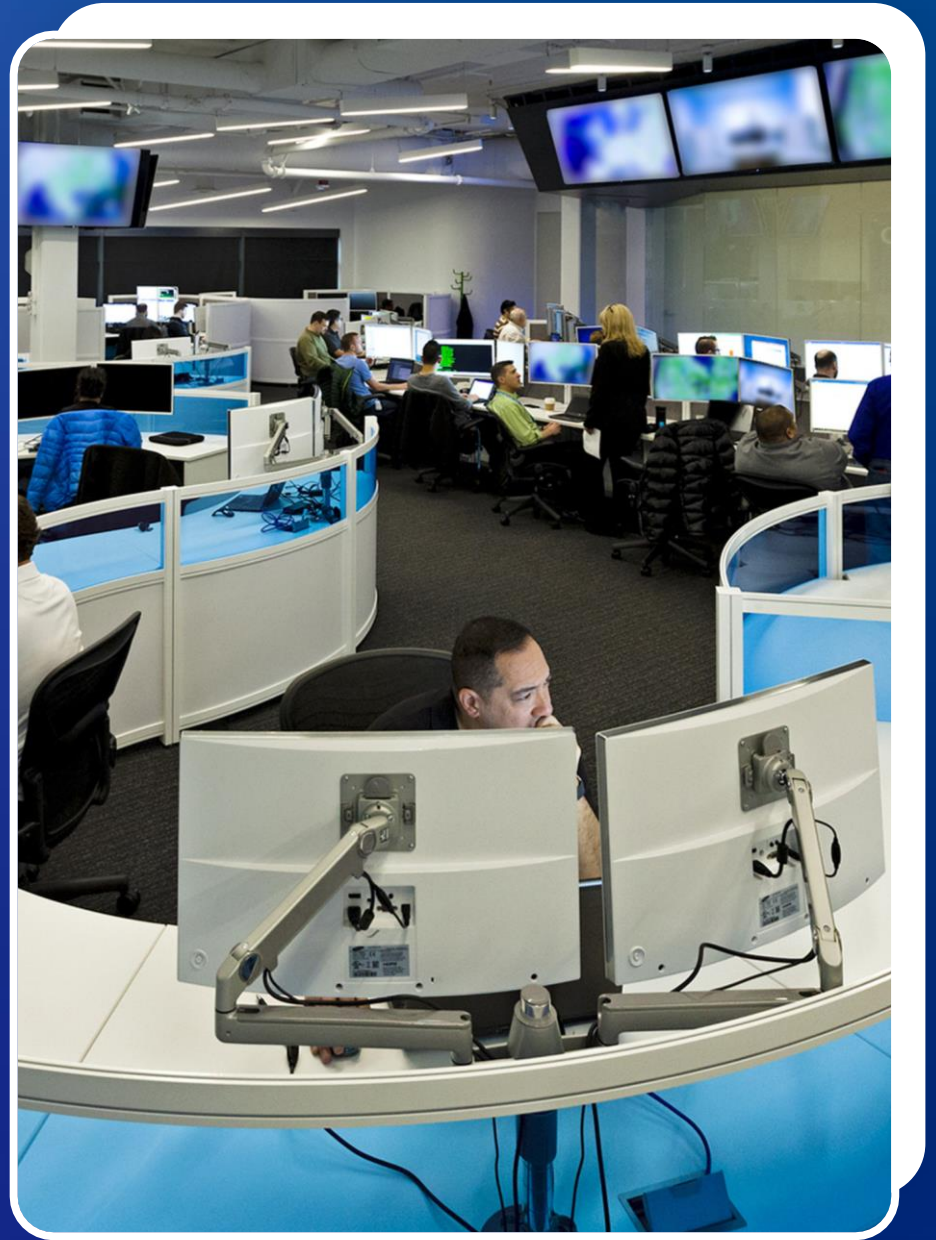


Adversarial
and Benign

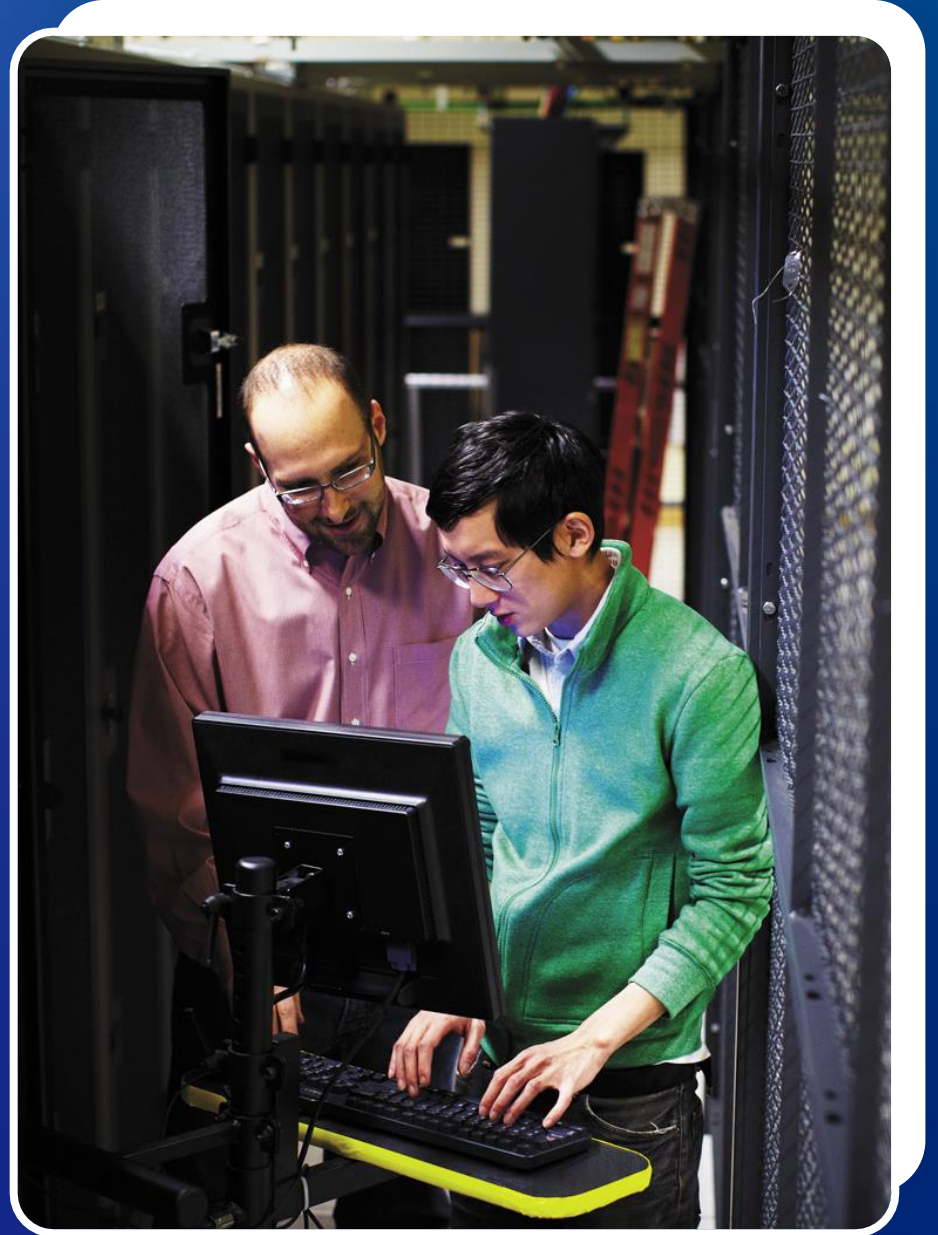


Rapidly Evolving
Tools and processes

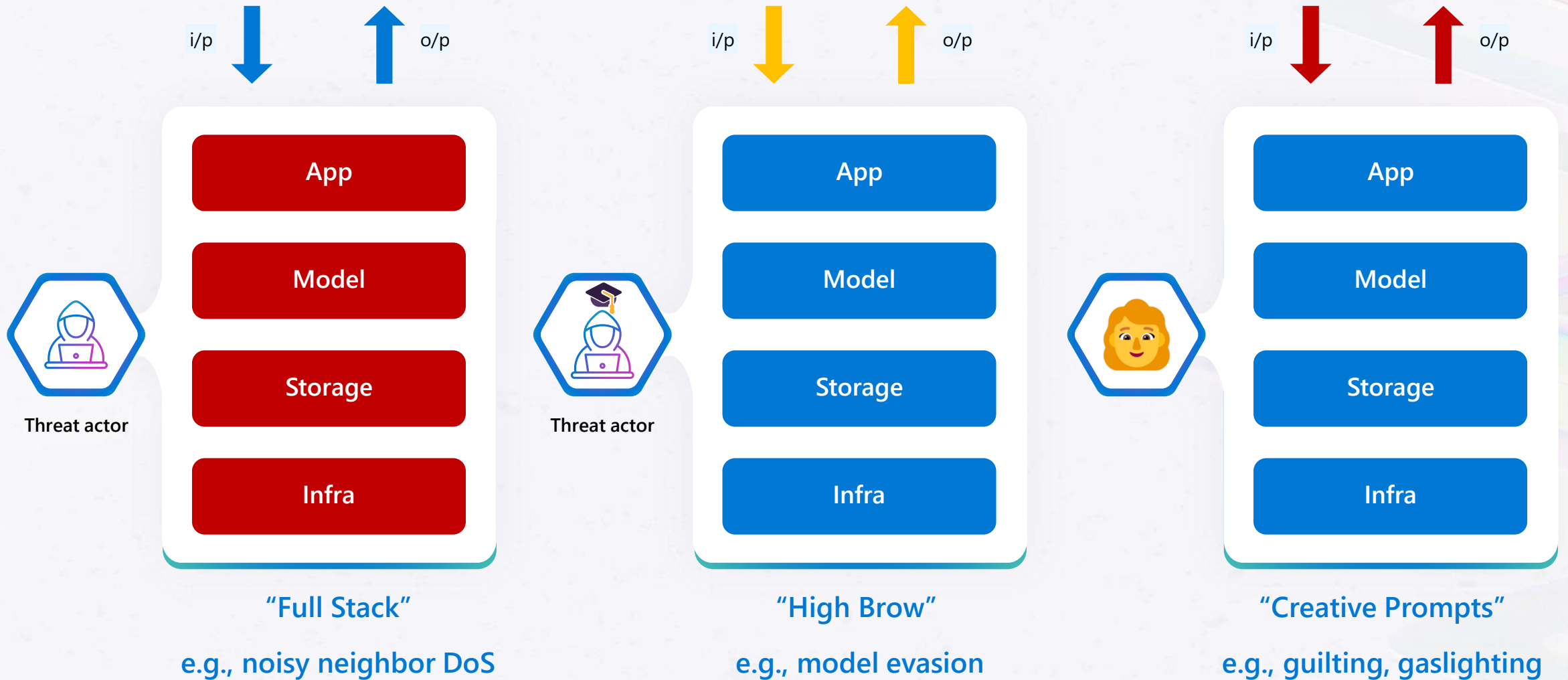
AI Red Teaming combines
the best of both worlds



AI Red Team = Probing for
Security + Responsible
AI Harms



Three Flavors of AI Red Teaming



Three Flavors of AI Red Teaming



Threat Actor

Full Stack

Focusing on the entire AI stack

Leveraging Traditional Security skills

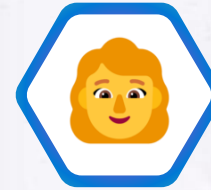


Threat Actor

High Brow

Focus only on the i/p and o/p

Leveraging Adversarial ML skills



Creative Prompt

Focuses on the i/p and o/p

Leverages a broad skillset to cause failures

AI-specific red teaming hardens the effectiveness of security protections



Expands the definition
and scope for AI



Focuses on failures
from both malicious
and benign personas

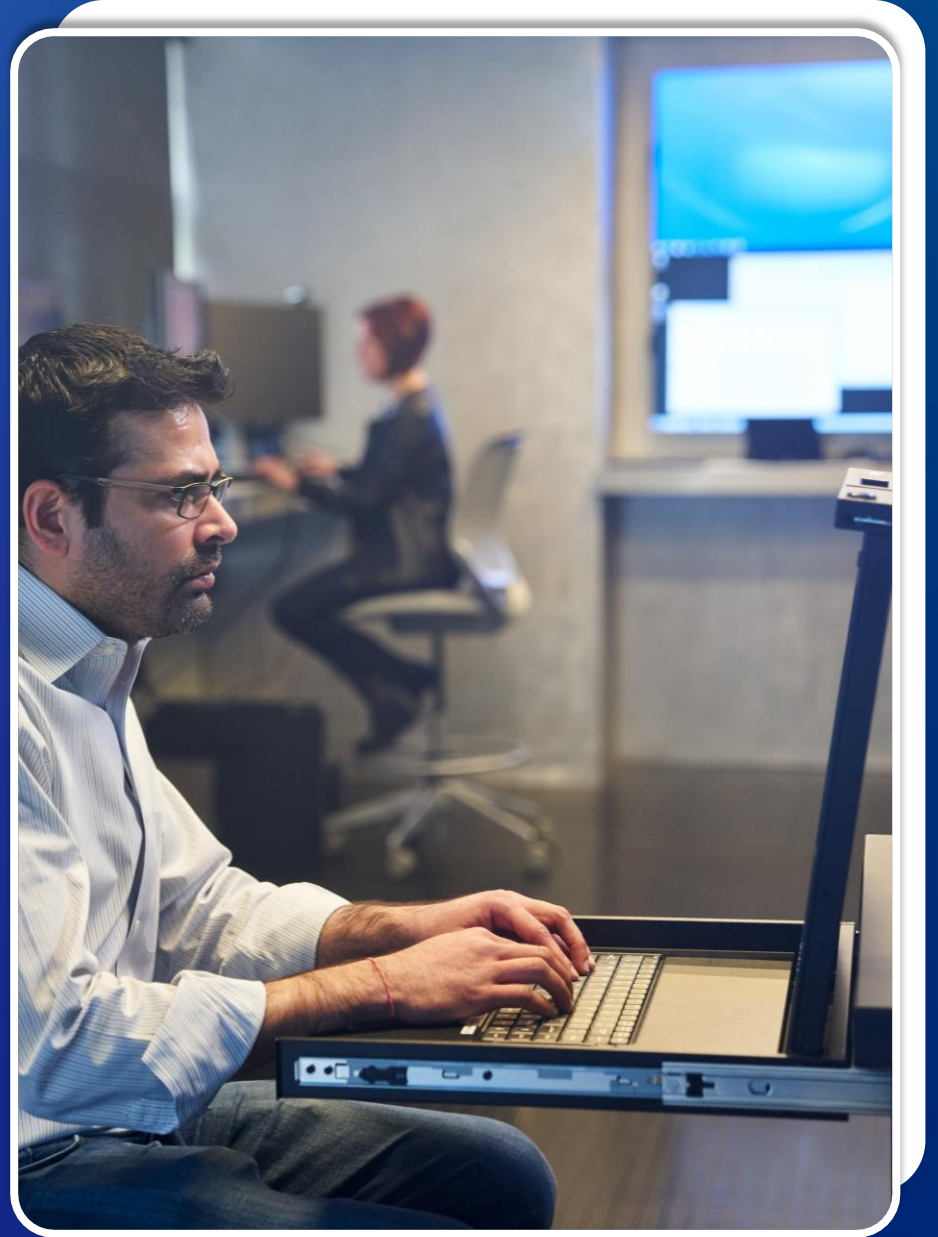


Recognizes that
AI systems are
constantly evolving

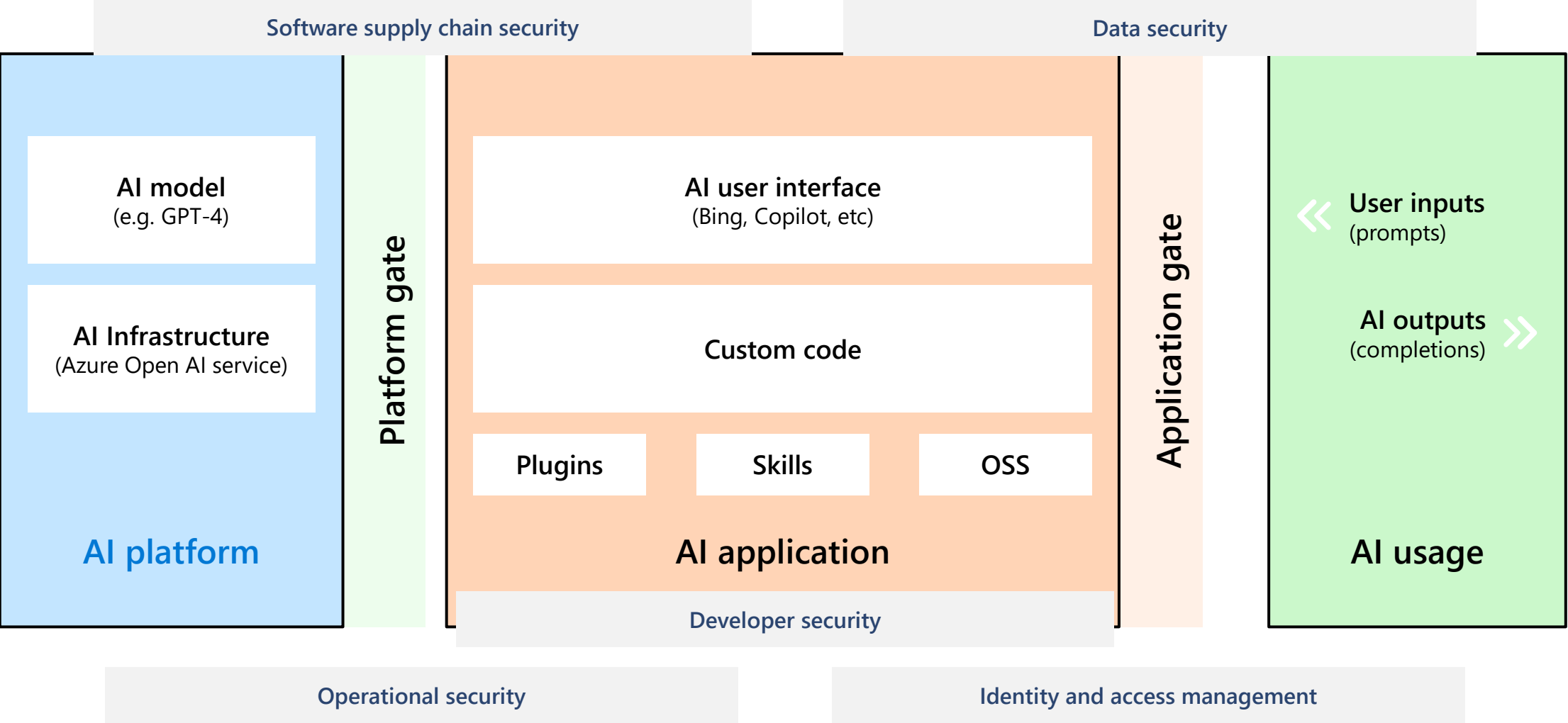
Learn More

<https://www.microsoft.com/en-us/security/blog/2023/08/07/microsoft-ai-red-team-building-future-of-safer-ai/>

Security controls for developers building AI-enabled applications



Security controls within AI systems



Threat Modelling Scenarios

MSRC AI Bug Bar

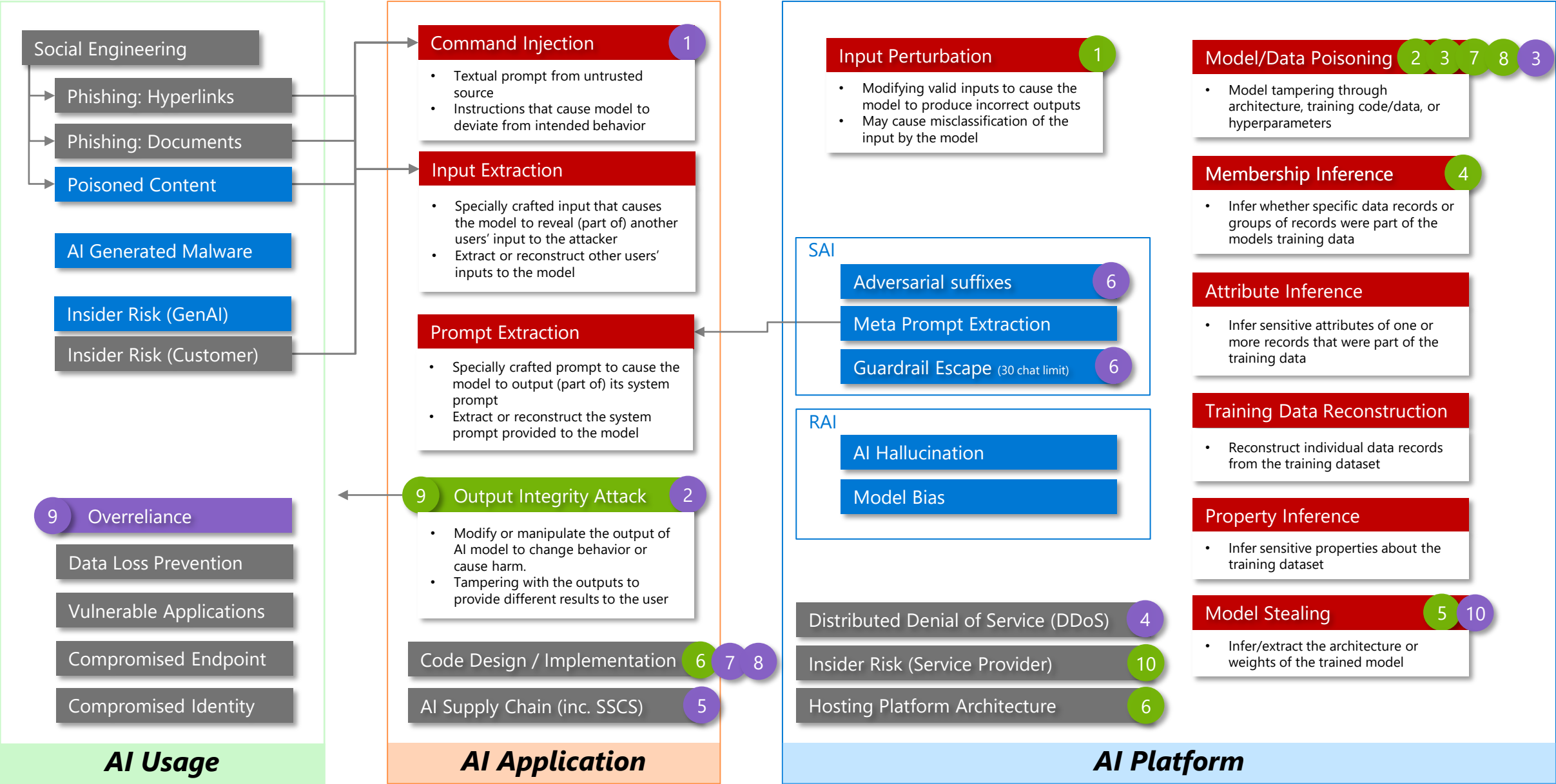
OWASP Top 10 for ML

Generative AI Specific

MITRE ATLAS

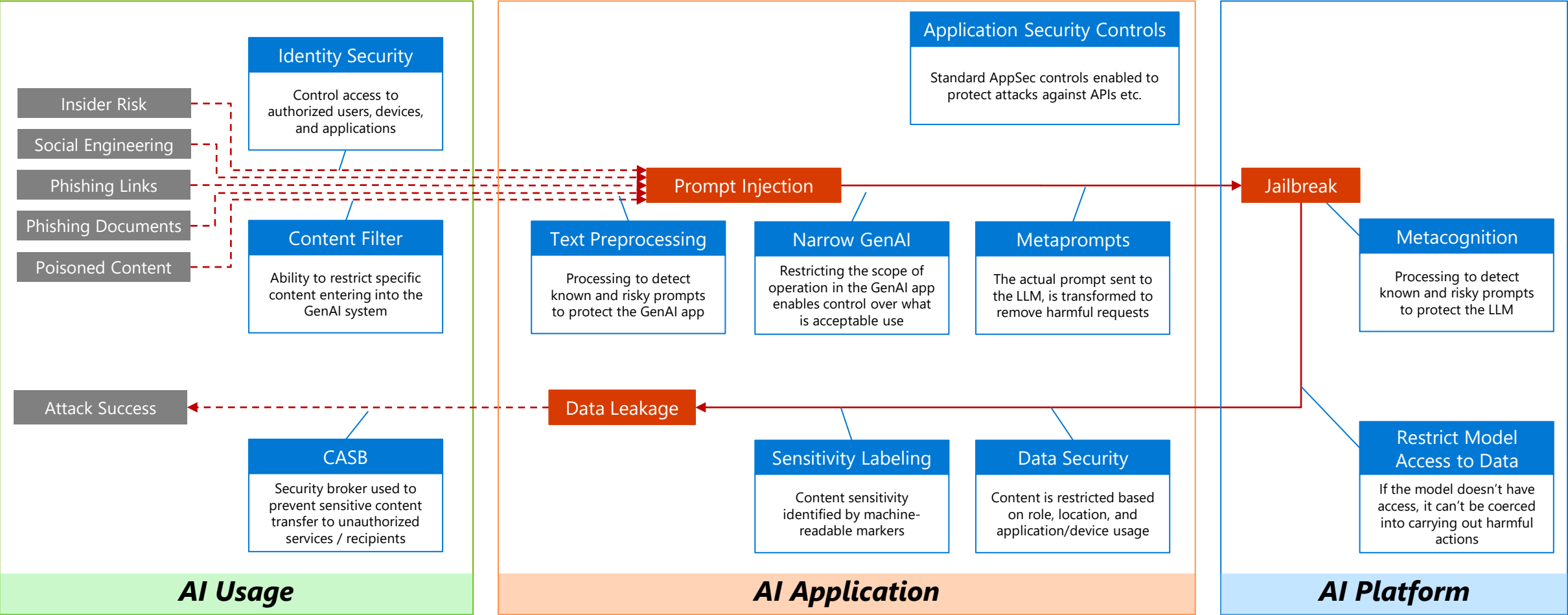
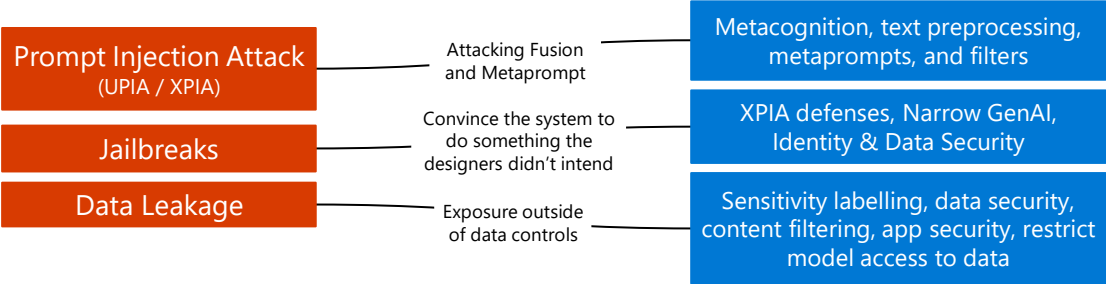
OWASP Top 10 for LLM

Common Cyber Threats



Threat Mapping Template

This framework provides a repeatable method of articulating both the vulnerabilities (red) and the mitigations (blue)



End-to-End Secure AI

SaaS (Copilot)

User Training

Customer needs to update their internal training to help users become productive with AI

Accountability

The user needs to remain accountable for the input and use of the output with AI

AI Acceptable Use Policy

Developing an AI Acceptable Use Policy is important for the identification of bad behavior

Administrative Controls

Configuration must be secure by default, with clear guidance to administrators for changes

Identity & Access Management

Built-in capabilities ensure customers can enable high-security controls

Data Governance

Copilot solutions honor the data governance of the existing solution they are created for

AI Usage

AI Plugins

Extending functionality via plugins must be carefully considered to protect integrity

Data Connectors

Internal and external data connectors must be carefully considered to protect data

Application Design & Implementation

Copilot solutions are developed following strict SDL governance with a deployment safety review board and AI Red Team

Application Infrastructure

Copilot solutions are made highly available, secure, and monitored from code to cloud

Application Safety Systems

Copilot specific safety systems are used to add additional layers of defense for AI

AI Application

Model Safety & Security Systems

The copilot inherits the safety and security systems built to protect the LLM

Model Accountability

Microsoft offers the pre-trained LLM capabilities as a private hosting

Model Tuning

For copilot solutions, models are not tuned, instead we use semantics and grounding

Model Design & Implementation

Microsoft builds, implements, and support the entire copilot stack end to end

Model Training Data Governance

Copilot runs on pre-trained LLM. Microsoft does not access the model or training data.

AI Compute Infrastructure

Azure has the largest ... (GPUs?)

AI Platform

		IaaS (BYO Model)	PaaS (Azure AI)	SaaS (Copilot)
AI Usage	User Training and Accountability			
	Usage Policy, Admin Controls			
	Identity, Device, and Access Management			
	Data Governance			
AI Application	AI Plugins and Data Connections			
	Application Design and Implementation			
	Application Infrastructure			
	Application Safety Systems			
AI Platform	Model Safety & Security Systems			
	Model Accountability			
	Model Tuning			
	Model Design & Implementation			
	Model Training Data Governance			
	AI Compute Infrastructure			

Microsoft

Model Dependent

Shared

Customer

Wrap up



Security for AI is an **ever-evolving process**



Develop

Continually evolving
secure by design
and by default
requirements



Test

Continual assessment
and AI-specific
red teaming



Monitor/Respond

Analyze 65T+ threat
signals, respond
to incidents



Evolve

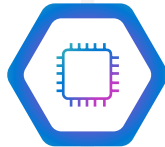
Learn, share,
collaborate



Further reading



[Microsoft Security Copilot documentation | Microsoft Learn](#)



[AI shared responsibility model – Microsoft Azure | Microsoft Learn](#)



[Best practices for AI security risk management | Microsoft Security Blog](#)



[Microsoft Cybersecurity Reference Architectures \(MCRA\) – Security documentation | Microsoft Learn](#)

