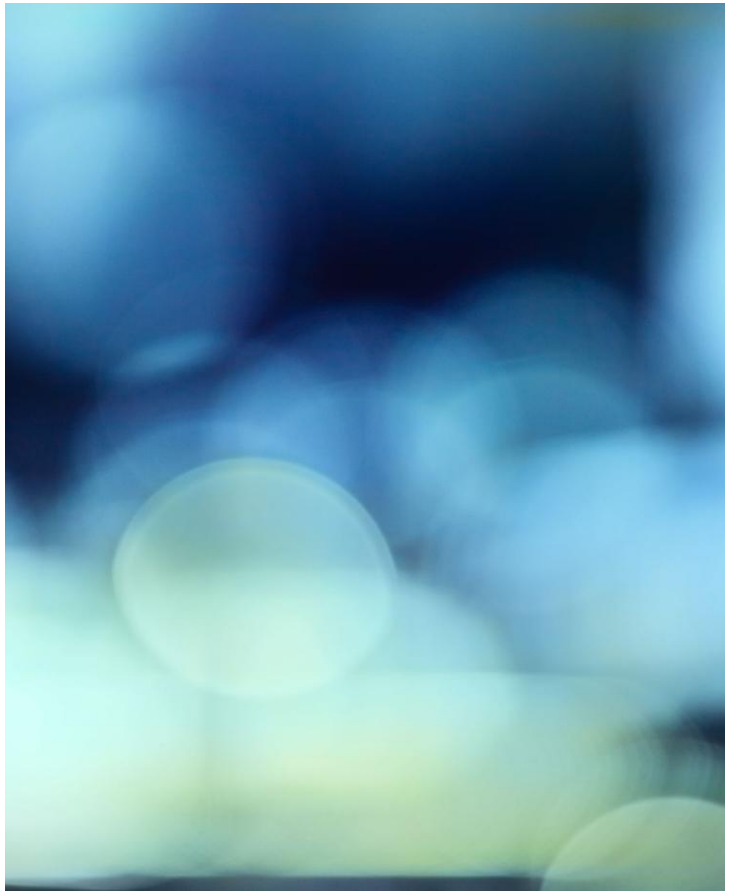
Securing AI – Notes from the field

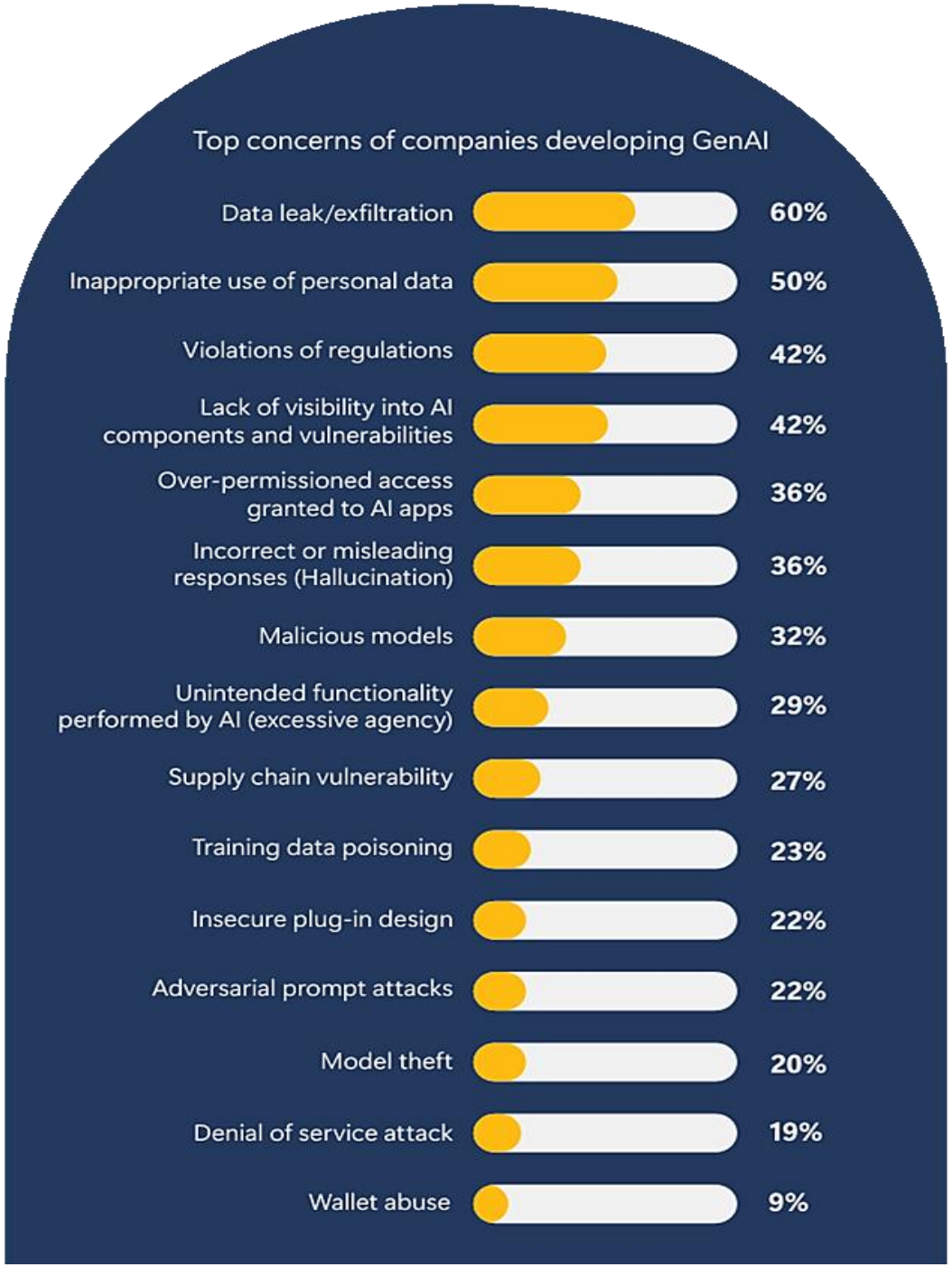
What Attendees Will Learn:

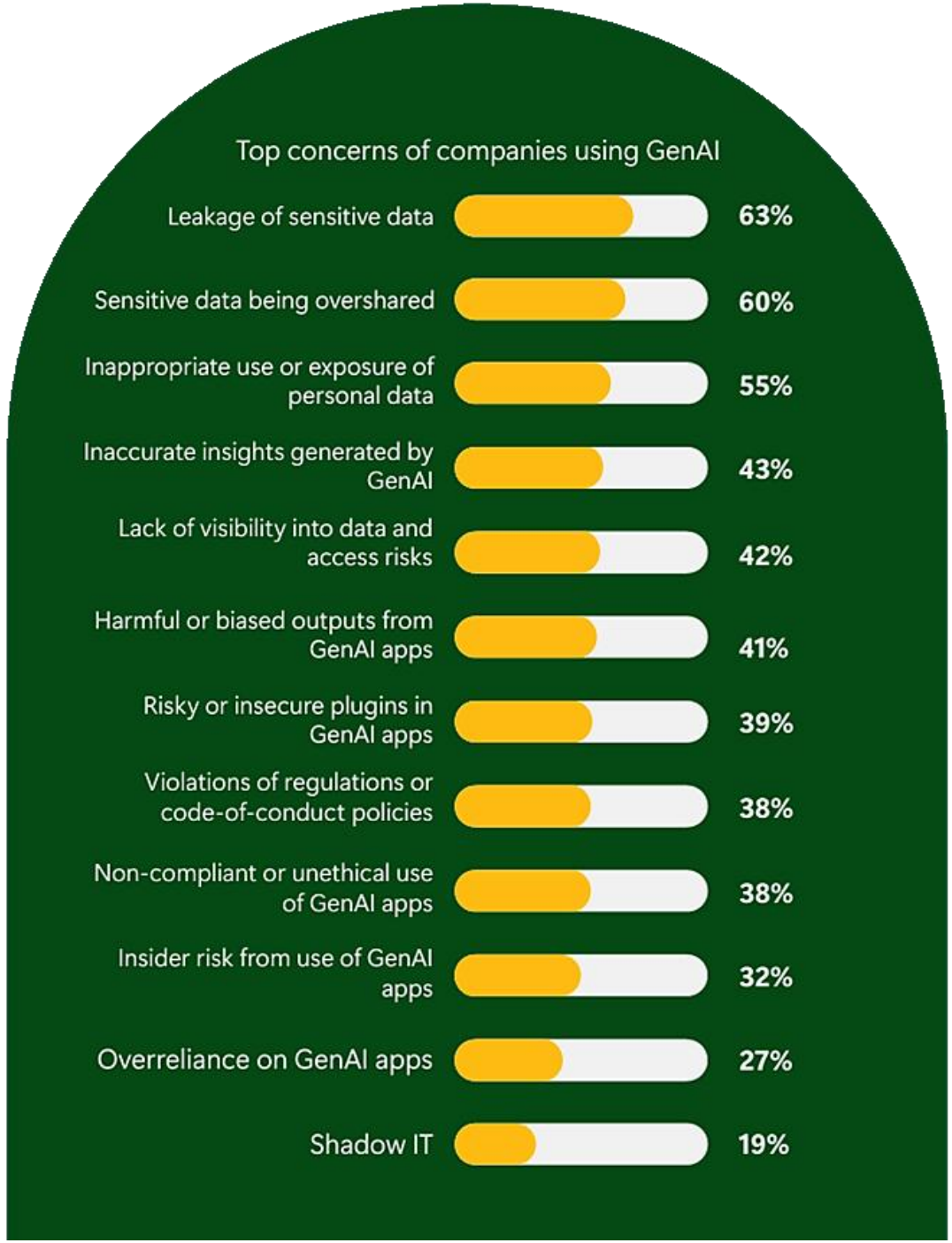
Identify how AI introduces new fraud attack vectors — including deepfake impersonation, AI-generated phishing, prompt injection attacks, and sensitive data disclosure — and recognize the warning signs in enterprise environments.

Evaluate the security gaps and common antipatterns that leave organizations vulnerable to AI-enabled fraud, including neglecting data foundations, applying only traditional security controls to dynamic AI systems, and insufficient access governance.

Apply a practical AI security framework — encompassing Zero Trust principles, the OWASP Top 10 for LLM Applications, and regulatory compliance requirements (GDPR, EU AI Act, NIST AI RMF, ISO 42001) — to strengthen fraud prevention, detection, and response capabilities within their organizations.




Top concerns of companies developing GenAI
Data leak/exfiltration 60%
Inappropriate use of personal data 50%
Violations of regulations 42%
Lack of visibility into AI components and vulnerabilities 42%
Over-permissioned access granted to AI apps 36%
Incorrect or misleading responses (Hallucination) 36%
Malicious models 32%
Unintended functionality performed by AI (excessive agency) 29%
Supply chain vulnerability 27%
Training data poisoning 23%
Insecure plug-in design 22%
Adversarial prompt attacks 22%
Model theft 20%
Denial of service attack 19%
Wallet abuse 9%

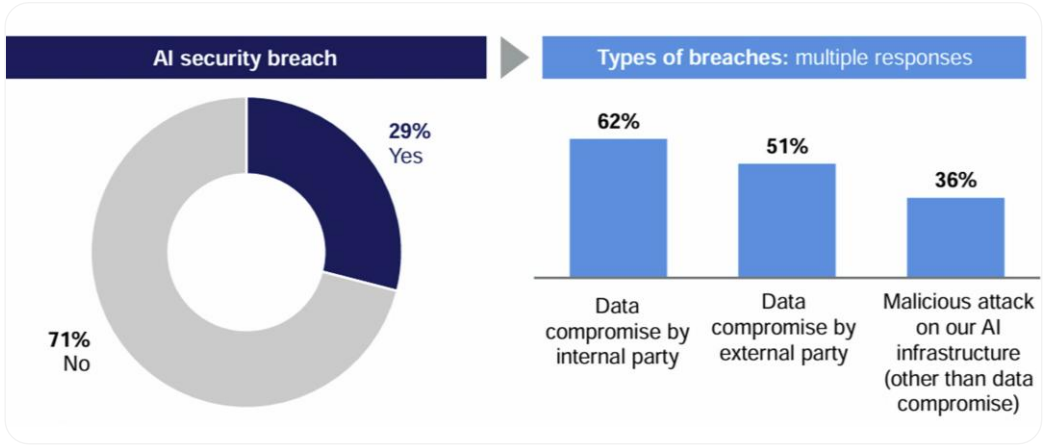

Top concerns of companies using GenAI
Leakage of sensitive data 63%
Sensitive data being overshared 60%
Inappropriate use or exposure of personal data 55%
Inaccurate insights generated by GenAI 43%
Lack of visibility into data and access risks 42%
Harmful or biased outputs from GenAI apps 41%
Risky or insecure plugins in GenAI apps 39%
Violations of regulations or code-of-conduct policies 38%
Non-compliant or unethical use of GenAI apps 38%
Insider risk from use of GenAI apps 32%
Overreliance on GenAI apps 27%
Shadow IT 19%

New Risks and Threats

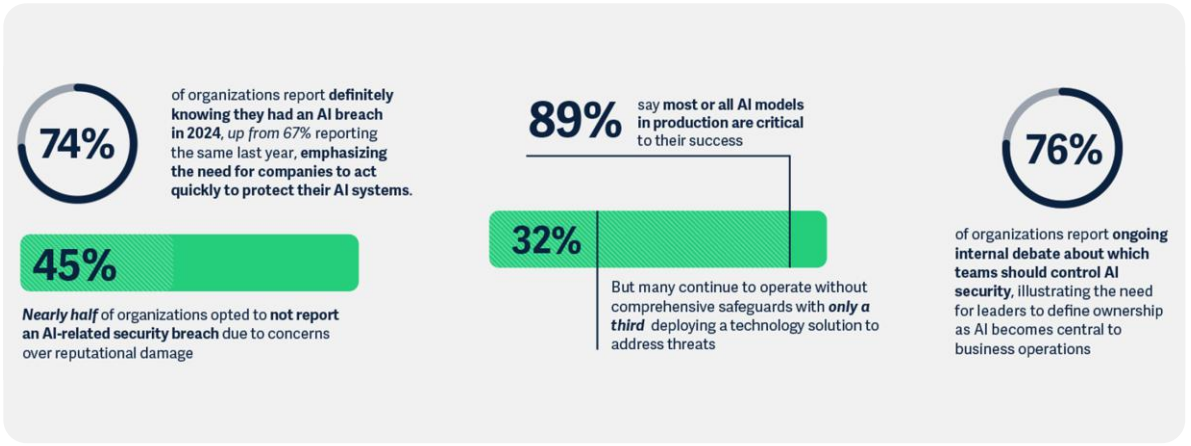
Addressing the evolving threat landscape is crucial to enabling trustworthy AI. GenAI introduces new attack surfaces, such as prompts, responses, training data, retrieval-augmented generation data, and models, effectively changing the risk landscape. In addition to managing traditional threat vectors, security and risk leaders also need to address amplified risks such as data leakage and data oversharing, and new risks such as prompt injections, hallucinations, and model vulnerabilities.



AI Security Trends



2023 Gartner AI in the Enterprise Survey



HIddenLayer 2024: HiddenLayer AI Threat Landscape Report Reveals AI Breaches on the Rise

97%

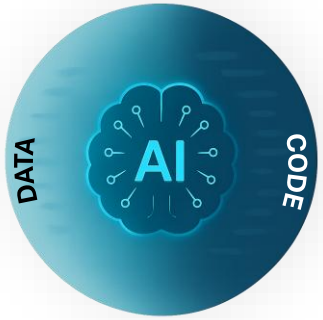
of organizations reported security incidents related to GenAI in the past year.

Capgemini “New Defenses, New threats” report Nov 2024: CRI_AI-and-gen-AI-in-cybersecurity_15112024.pdf



2024 Microsoft Guide for Securing the AI-Powered Enterprise | Security Insider

AI Requires New Security Measures (1 of 2)



AI blurs line between data & code/instructions

Increases value *and* risk of abuse/attack

Classic Application **Components**

AI Components

Predictable Logic

Consistent (deterministic) outcomes based on execution of classic programming (and classic bug types ☺)

Dynamic Logic

Pattern of variable outcomes based on model design, training data, real-time inputs, other controls

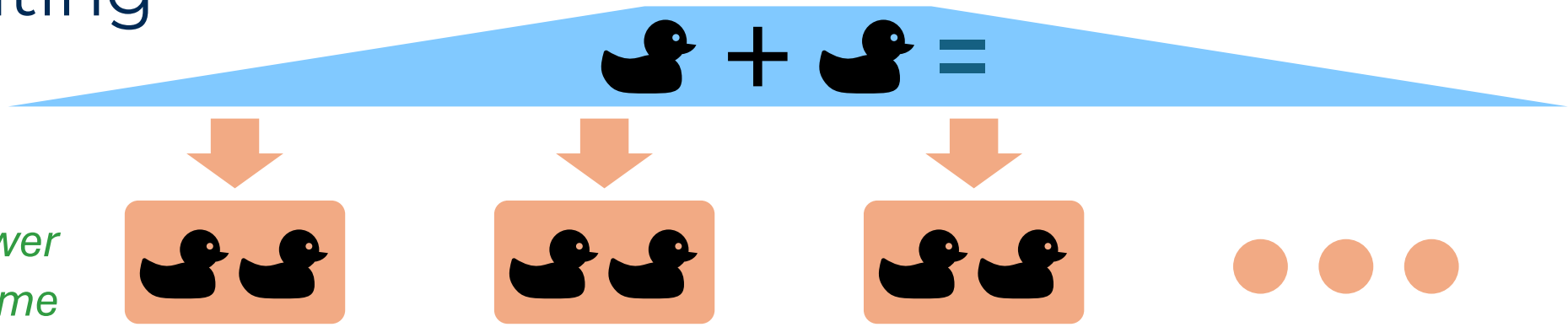
Different Technical Exploits & Defenses for

Precise Interruption/ Redirection of Logic Flow

General biases & hallucinations in outcomes

Classic computing

is deterministic

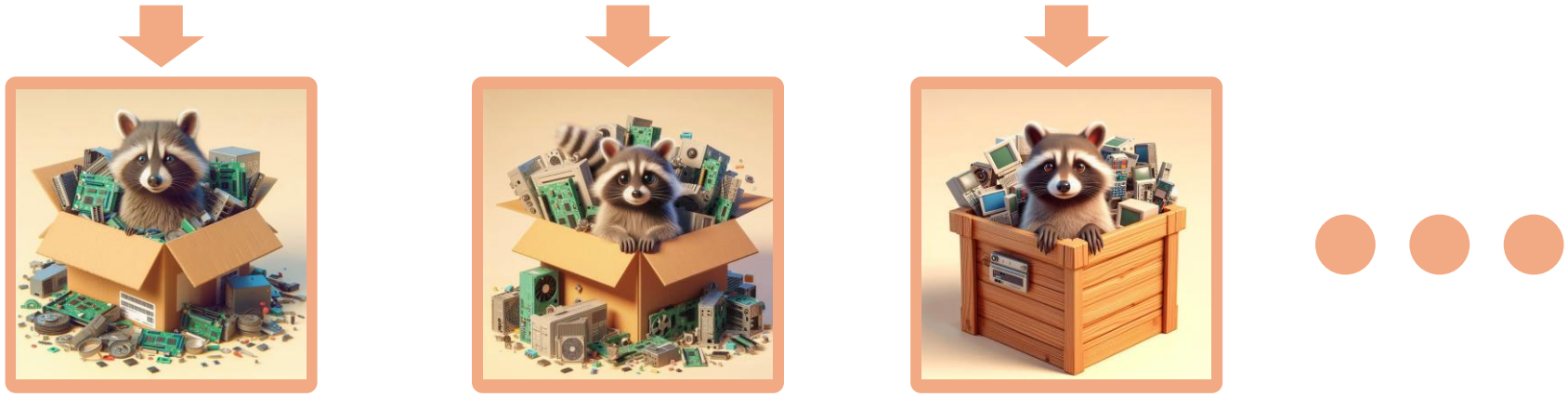


Same exact answer every time

Generative AI

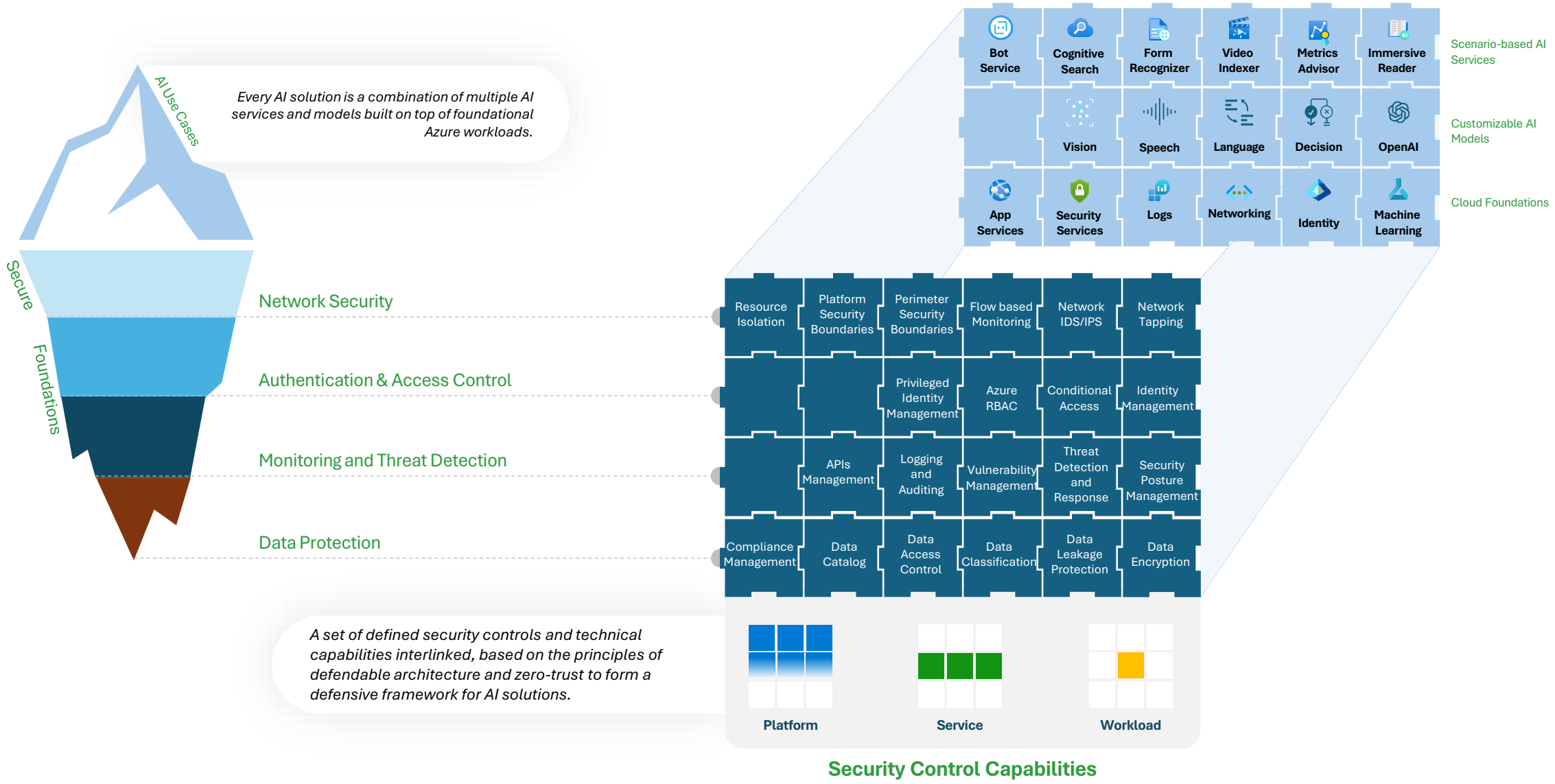
is dynamic

"A cute raccoon peeks out of a box filed with recycled computers and computer parts on a cream background in a 3D hyper-surreal style."



Similar but not identical results

AI Requires New Security Measures (2 of 2)



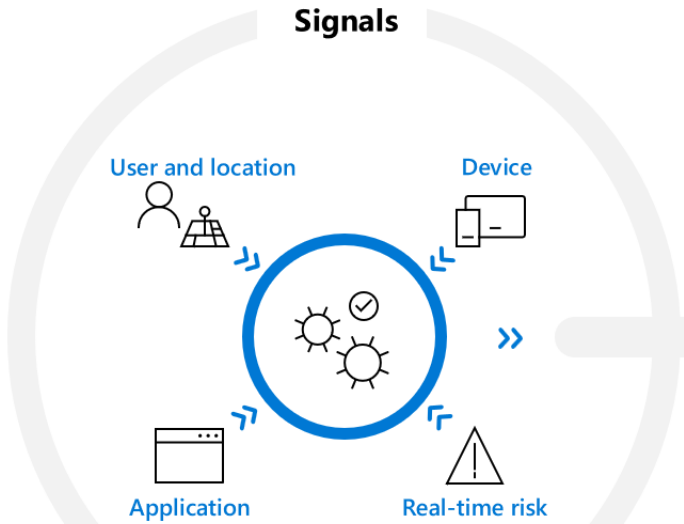
Enable Zero Trust with Strong Authentication, Adaptive Policies and multi-layers protection

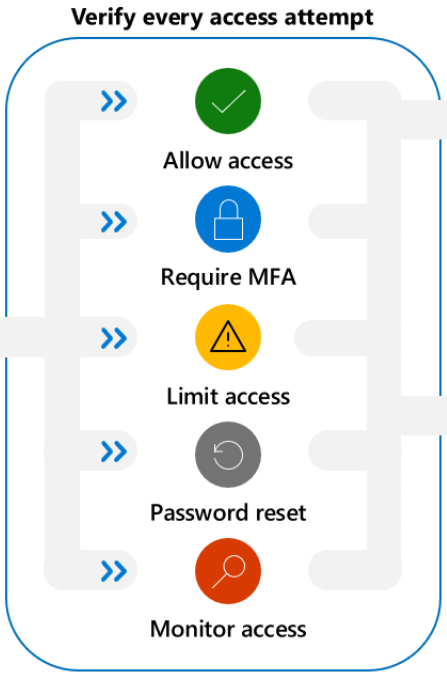
Hybrid and Multi-clouds Identity Perimeter

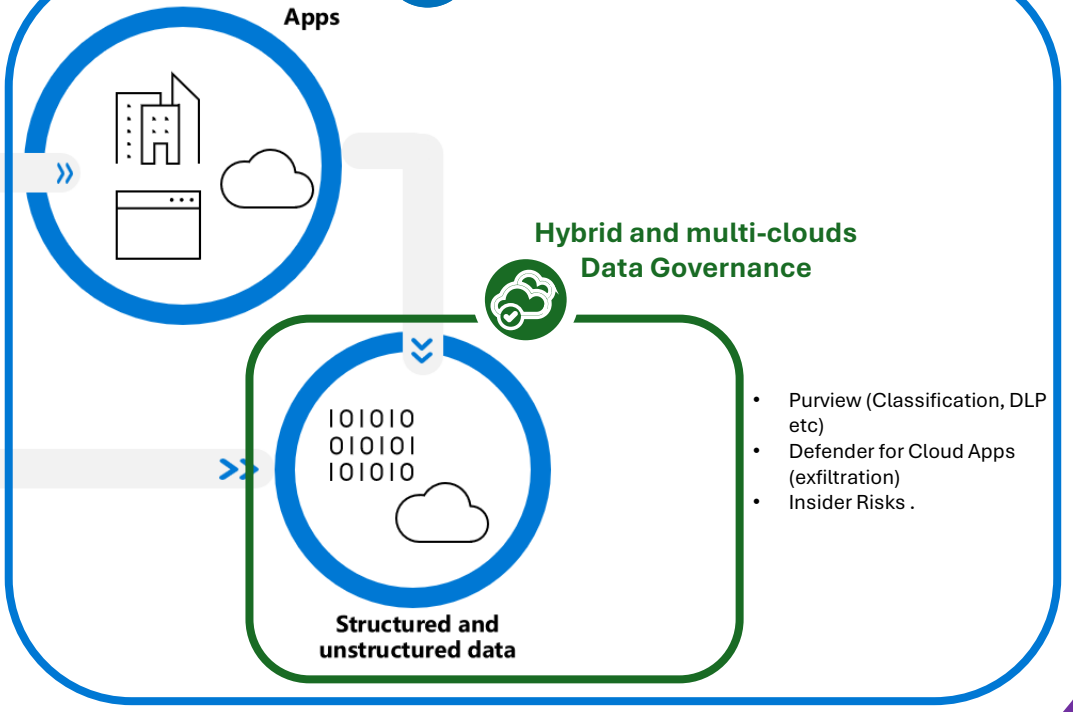
- Risk-based Conditional Access Policies
- App Proxy pre-Auth (legacy)
- Defender for Coud Apps (session validation - CAEP)

Hybrid and multi-clouds Access Management

- RBAC / ABAC
- IGA: Access Packages, Access review, Approval
- Permissions Manager







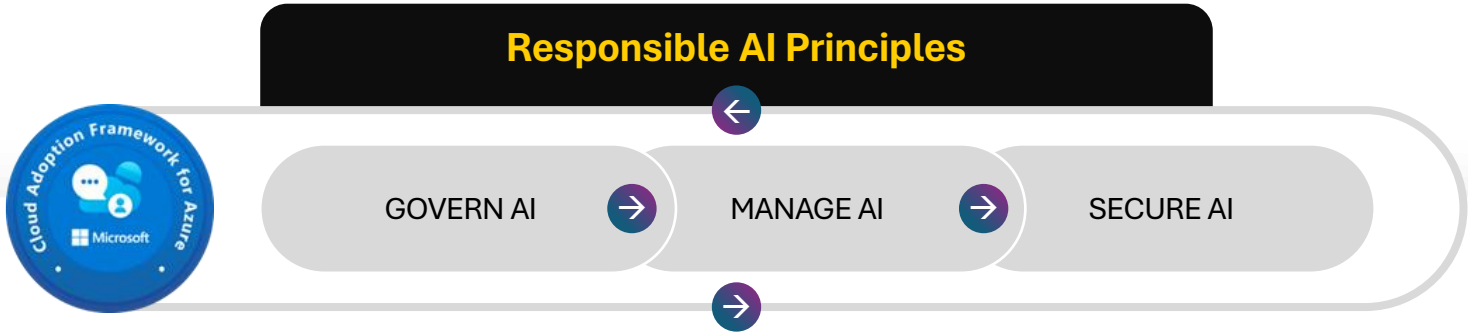
- Purview (Classification, DLP etc)
- Defender for Cloud Apps (exfiltration)
- Insider Risks .

Security Team(s)

IAM Team

App & Data owners

AI Security Framework



Discover AI Security Risks

- *Identify AI security risks*
- *Identify AI data risks*
- *Identify AI model risks*
- *Conduct red team testing*
- *Conduct periodic risk assessments*

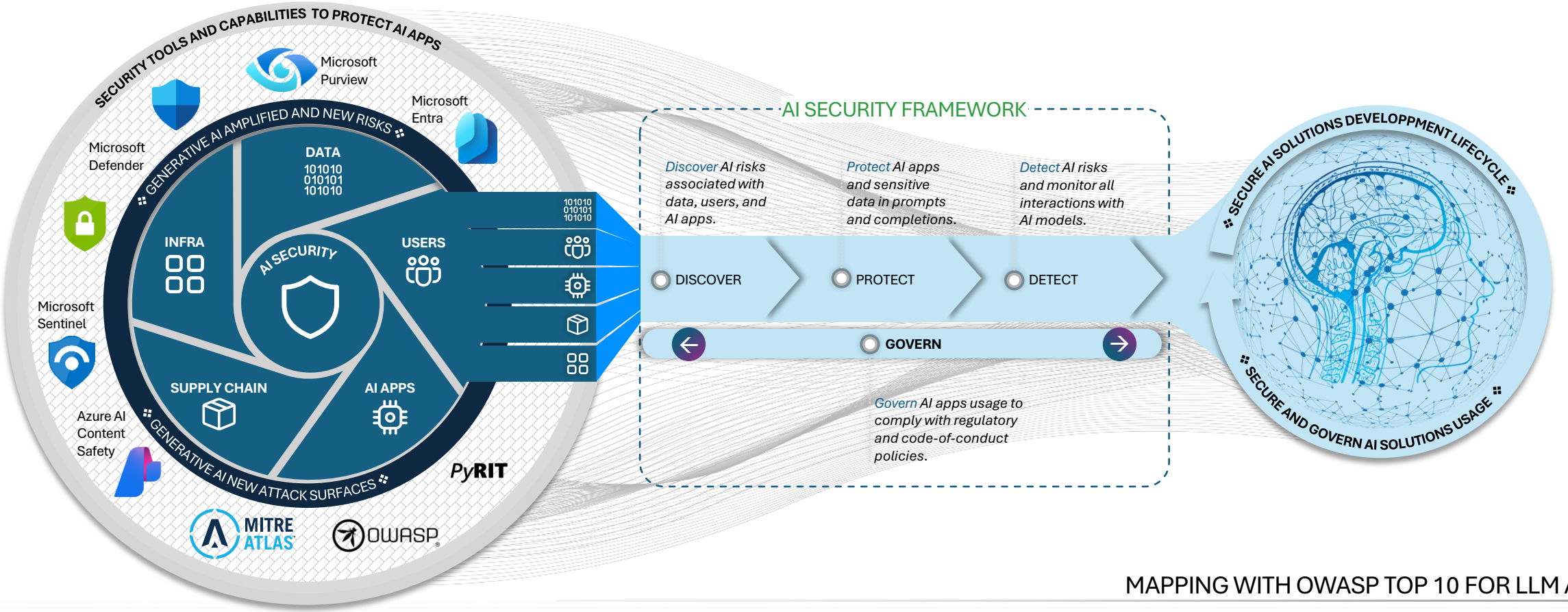
Protect AI Resources and Data

- *Centralize AI asset inventory*
- *Secure Azure AI platforms*
- *Apply workload security controls*
- *Control data access*
- *Safeguard sensitive data*

Detect AI Security Threats

- *Detect AI risks*
- *Detect AI security incidents*
- *Apply workload-specific monitoring*

Secure AI Usage with Microsoft Security



MAPPING WITH OWASP TOP 10 FOR LLM APPS

LLM01
Prompt Injection
This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injection overwrite system prompts, while indirect ones manipulate inputs from external sources.

LLM02
Insecure Output Handling
This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

LLM03
Training Data Poisoning
Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.

LLM04
Model Denial of Service
Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

LLM05
Supply Chain Vulnerabilities
LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins add vulnerabilities.

LLM06
Sensitive Information Disclosure
LLM's may inadvertently reveal confidential data in its responses, leading to unauthorized data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.

LLM07
Insecure Plugin Design
LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in severe consequences like remote code execution.

LLM08
Excessive Agency
LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

LLM09
Overreliance
Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

LLM10
Model Theft
This involves unauthorized access, copying, or exfiltration of proprietary LLM models. The impact includes economic losses, compromised competitive advantage, and potential access to sensitive information.

AI Security Compliance and Regulations

			Key Focus/Requirements	Impact on AI Security Compliance	Examples of Adoption/ Implementation
GDPR EU General Data Protection Regulation	**REGULATION (Data Privacy Law)**	***European Union (global reach)***	**Personal data protection** – Requires user consent for data use, transparency in automated decisions, strong data security, and respect for individual data rights.	Organizations must ensure AI systems handling personal data comply with privacy rules; non-compliance can lead to fines up to **€20 Million or 4%** of global annual revenue.	Widely adopted as a global privacy standard (any company processing EU citizens' data must comply). Example: Privacy adaptations in AI chatbots – e.g., OpenAI adjusted ChatGPT's data handling after Italy's GDPR inquiry.
EU Artificial Intelligence Act	**REGULATION (AI-specific Law)**	***European Union***	**Risk-based AI governance** – Prohibits dangerous AI uses (e.g. manipulative AI, social scoring); imposes strict requirements on "high-risk" AI (e.g. bias testing, documentation, human oversight) before deployment.	Will mandate robust AI oversight and audits; **non-compliance can incur fines up to €35 Million or 7%** of global turnover. Sets a new compliance bar globally for AI products entering the EU market.	Taking effect in phases (2025–2026); companies are preparing now (especially in finance, healthcare). Example: European banks are assessing their AI credit models to meet high-risk system criteria ahead of enforcement.
NIST AI Risk Management Framework	**FRAMEWORK (Guidance)**	***United States (International use)***	**Voluntary AI risk management** guidelines – Provides a structured approach with core functions Govern, Map, Measure, Manage to embed trustworthiness (e.g. fairness, transparency, security) into AI design and use.	Not legally mandatory, but widely influential. Helps organizations improve AI **safety and compliance posture** by aligning with best practices. Often used as a blueprint to meet or anticipate regulatory requirements (e.g., U.S. agencies and many firms adopt NIST's framework to manage AI risks).	Embraced by industry and government. Example: U.S. federal agencies are directed to use NIST's AI RMF for AI projects, and companies like Microsoft align their Responsible AI programs with NIST's categories.
ISO 42001/2023	**CERTIFICATION (ISO Standard)**	***International***	**AI governance & oversight** – A certifiable **AI Management System (AIMS)** standard (published Dec 2023) that requires organizations to implement controls for AI ethics, risk assessment, data protection, and security throughout the AI lifecycle. Aligns with global AI principles (e.g., transparency, accountability) to ensure **responsible AI** operations.	Companies can undergo independent audit to **certify compliance** with this standard, demonstrating they effectively manage AI risks and adhere to best practices. Supports meeting regulatory expectations (it's aligned with the EU AI Act and NIST RMF), giving stakeholders assurance of trustworthy AI.	Early adoption underway. Example: Microsoft's AI programs (e.g. Microsoft 365 Copilot) achieved ISO 42001 certification in 2025, providing external validation of its AI risk management. Many firms are seeking certification to build trust with clients and regulators.

Business Impact and Benefits

Enable Responsible AI

Align AI development with leading frameworks and responsible AI principles.

Accelerate with Confidence

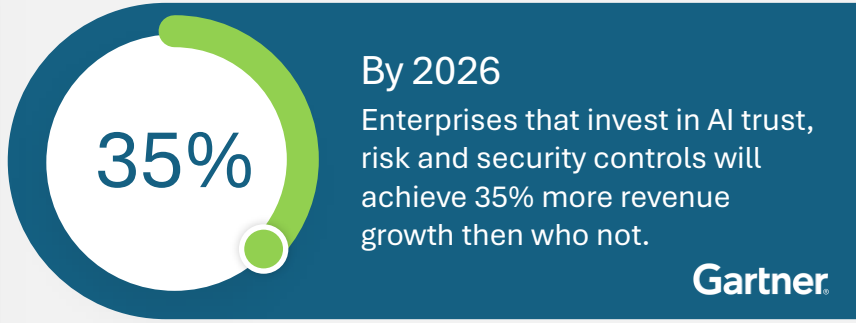
Innovate faster while managing AI-specific risks.

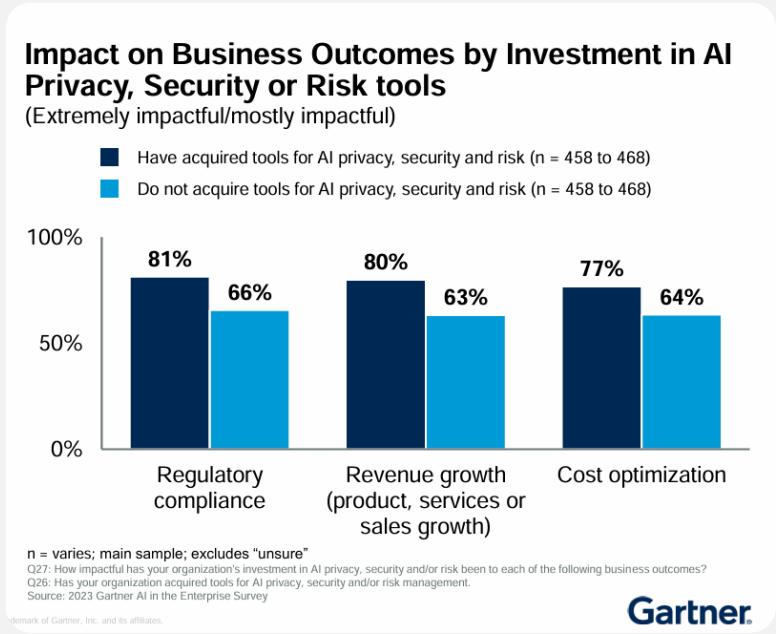
Reduce Exposure

Minimize the risk of data breaches, adversarial attacks, and regulatory penalties.

Build Trust

Demonstrate a proactive commitment to secure and ethical AI.





Thank you

