AI4Leaders
with passion and experience to success

GOhack25
Security in times of AI

Adam D. Wisniewski

08. November 2025

Painful learning process all over again ?

IBM

IBM **Newsroom** News Media resources Inside IBM Blog

IBM Report: 13% Of Organizations Reported Breaches Of AI Models Or Applications, 97% Of Which Reported Lacking Proper AI Access Controls

U.S. breach costs rise to $10.22 million, despite the global average cost of a breach decreasing to $4.44 million; Only 49% of breached organizations plan to invest in security

Jul 30, 2025

Forbes

BREAKING | BUSINESS

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

By Molly Bohannon, Former Staff. Molly Bohannon has been a Forbes news reporter since 2023.

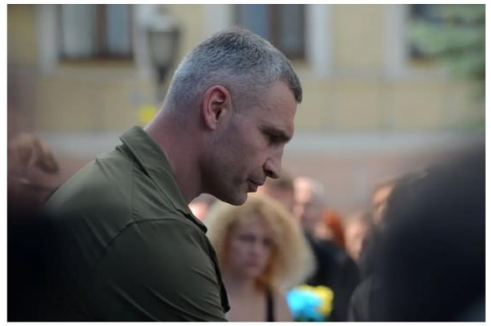
Published Jun 08, 2023 at 02:06pm EDT, Updated Jun 08, 2023 at 03:42pm EDT

International NZZ

«Moderne Kriegsführung»: Deep-Fake-Anrufe im Namen Klitschkos in Berlin, Wien und Madrid beunruhigen Europa

Per Video liess sich ein angeblicher Witali Klitschko in europäische Rathäuser schalten, konferierte über den Ukraine-Krieg. Die Täuschung sorgt für Unruhe.

25.06.2022, 18.50 Uhr min



Per Video liess sich ein angeblicher Vitali Klitschko in europäische Rathäuser schalten. Hier erweist der echte Klitschko, Kiews Bürgermeister, einem ukrainischen Kämpfer die letzte Ehre.

Evgen Kotenko / Imago

A REUTERS INVESTIGATION

We set out to craft the perfect phishing scam. Major AI chatbots were happy to help.

Reuters and a Harvard University researcher used top chatbots to plot a simulated phishing scam – from composing emails to tips on timing – and tested it on 108 elderly volunteers. The bots' persuasive performance shows how AI is arming criminals for industrial-scale fraud.

By STEVE STECKLOW and POPPY MCPHERSON | Filed Sept. 15, 2025, 10:30 a.m. GMT

Tay: Microsoft issues apology over racist chatbot fiasco

25 March 2016 · 385 Comments

The AI was taught to talk like a teenager

Dave Lee ›
North America technology reporter

Microsoft has apologised for creating an artificially intelligent chatbot that quickly turned into a holocaust-denying racist.

Forbes

BREAKING | BUSINESS

Samsung Bans ChatGPT Among Employees After Sensitive Code Leak

By Siladitya Ray, Forbes Staff. Siladitya Ray is a New Delhi-based Forbes news... Follow Author

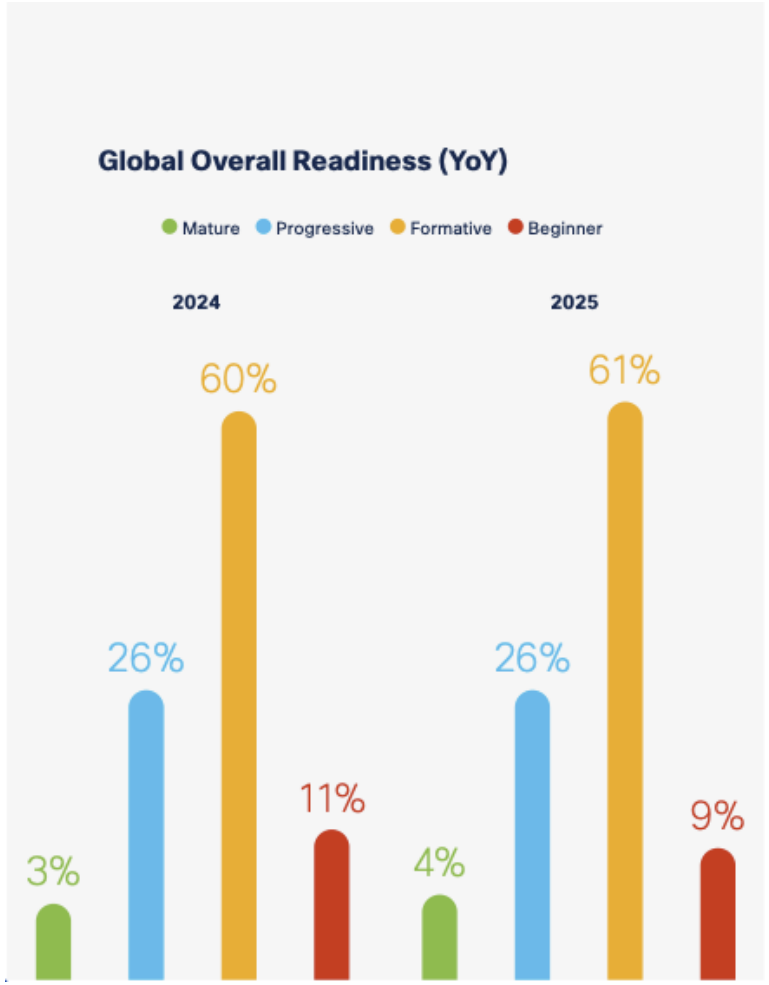
Published May 02, 2023 at 07:17am EDT, Updated May 02, 2023 at 07:31am EDT

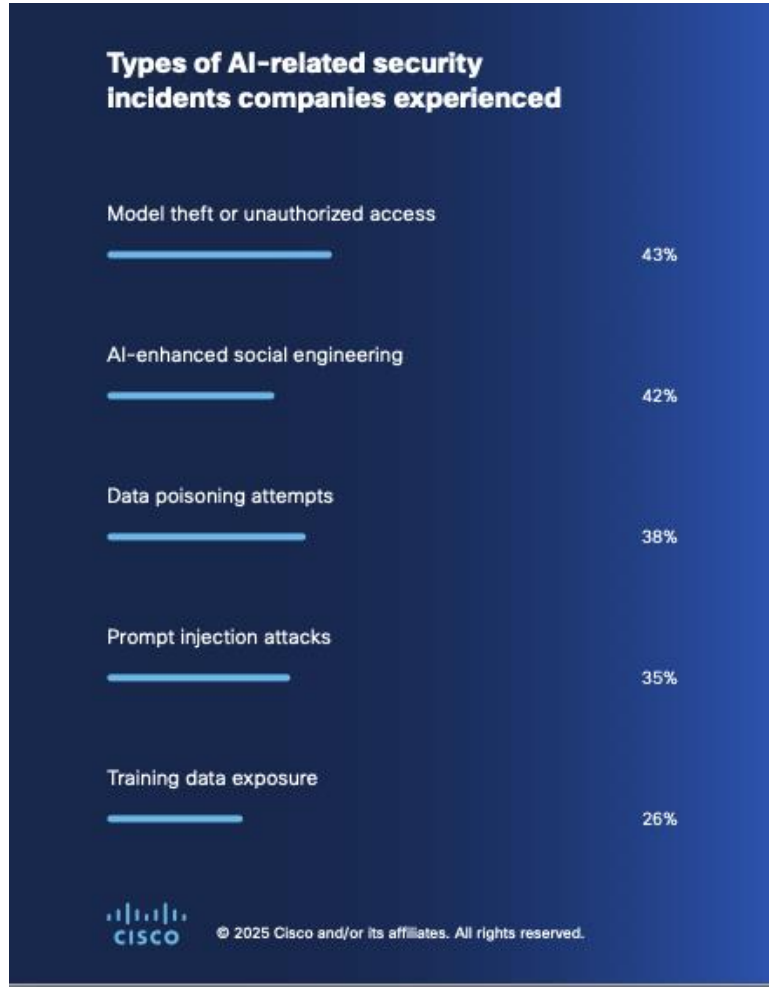
Where do we stand today ?

➡ Everyone has grown to appreciate AI in private, and doesn't want to be without at work.
+ When we get tangible value, we often forget all caution.

** PwC 2025 Global Artificial Intelligence Survey* and *IBM Global AI Adoption Index 2025*
*** Gartner AI Risk Survey 2025*
**** Forrester AI Trust Study 2025*

Common strategy – Let‘s wait and see..





Are new security measures really needed?

AI4Leaders

AI

Learns from data and **improves over time**.	**Learning and adaptation**	Static, requires reprogramming to make changes.
Makes **independent decisions** based on data.	**Autonomy in decision-making**	Follows predefined rules.
Handles complex, **unstructured tasks** such as image recognition.	**Complexity of the tasks**	Focuses on structured, rule-based tasks.
Works with large, unstructured data (e.g. text, images).	**Interaction with data**	Only processes structured data.
Imitates human thinking and creativity.	**Human intelligence**	Does not simulate human perception.
Generalised to new situations.	**Generalisation vs. specialisation**	Specialised in certain tasks.
Continually **evolves** with new data.	**Evolution over time**	Remains unchanged without updates.
Probabilistic, provides flexible predictions.	**Probabilistic vs. deterministic nature**	Deterministic, follows precise rules.
Requires considerable computing power and data.	**Resource intensity**	More resource-efficient.
Understands context and nuances (e.g. language).	**Contextual understanding**	Only understands the context if this has been explicitly programmed.

Traditional technologies

➡ Almost like a knowledgeable, but unexperienced employee

What is holding decision-takers back?

Use Outpaces Controls

Rapid AI adoption is evolving faster than traditional policy, security, and control cycles can adapt.

Historic Data Governance Gaps

Years of inconsistent data classification mean sensitive data is not "AI-ready" and can be easily exposed.

Regulatory Uncertainty

The legal landscape is still being formed, making it difficult to define a stable, long-term compliance target.

Over-Reliance on Providers

Trusting vendor claims without due diligence on their data handling and "black box" models creates third-party risk.

Widespread Knowledge Gaps

A lack of deep AI understanding across Management, IT, and users hinders the ability to manage new, complex risks.

Overestimation of Readiness

A tendency to overestimate existing AI knowledge at all levels leads to unidentified risks and a false sense of security.

- ➡ Management is often happy to «leave the matter» to the employees.
- ➡ Management is putting off this important but complex issue and waiting to see what happens.

What are common weaknesses?

Management

Category	Cause
Governance gap	No clear ownership of AI risk
Strategic mindset	Focus on speed over control
Financial constraint	Limited budgets & unclear ROI
External environment	Regulatory & legal uncertainty
Framework confusion	Fragmented standards
Skills & capability gap	Lack of AI/security know-how
Third-party dependency	Weak vendor ecosystem

Employees

Category	Cause
Behavioural risk	Shadow AI use
Knowledge gap	Low AI security literacy
Policy gap	No internal guidance or training

Risks today, and risks tomorrow - examples

Today

Category	Threat	Probability	Severity	Risk Level
External	AI-augmented phishing & deepfake social engineering	High	Major	Critical
External	Autonomous malicious AI agents (spam/fraud bots)	Medium	Major	High
External	Model/API key theft and supply-chain compromise	Medium	Major	High
Internal	Prompt injection & malicious content in RAG/chatbots	High	Major	Critical
Internal	Data leakage via RAG or vector stores	High	Major	Critical
Internal	Over-permissioned AI tools and agents	Medium–High	Major–Critical	Critical
Internal/ External	Data poisoning or model manipulation	Medium	Major	High

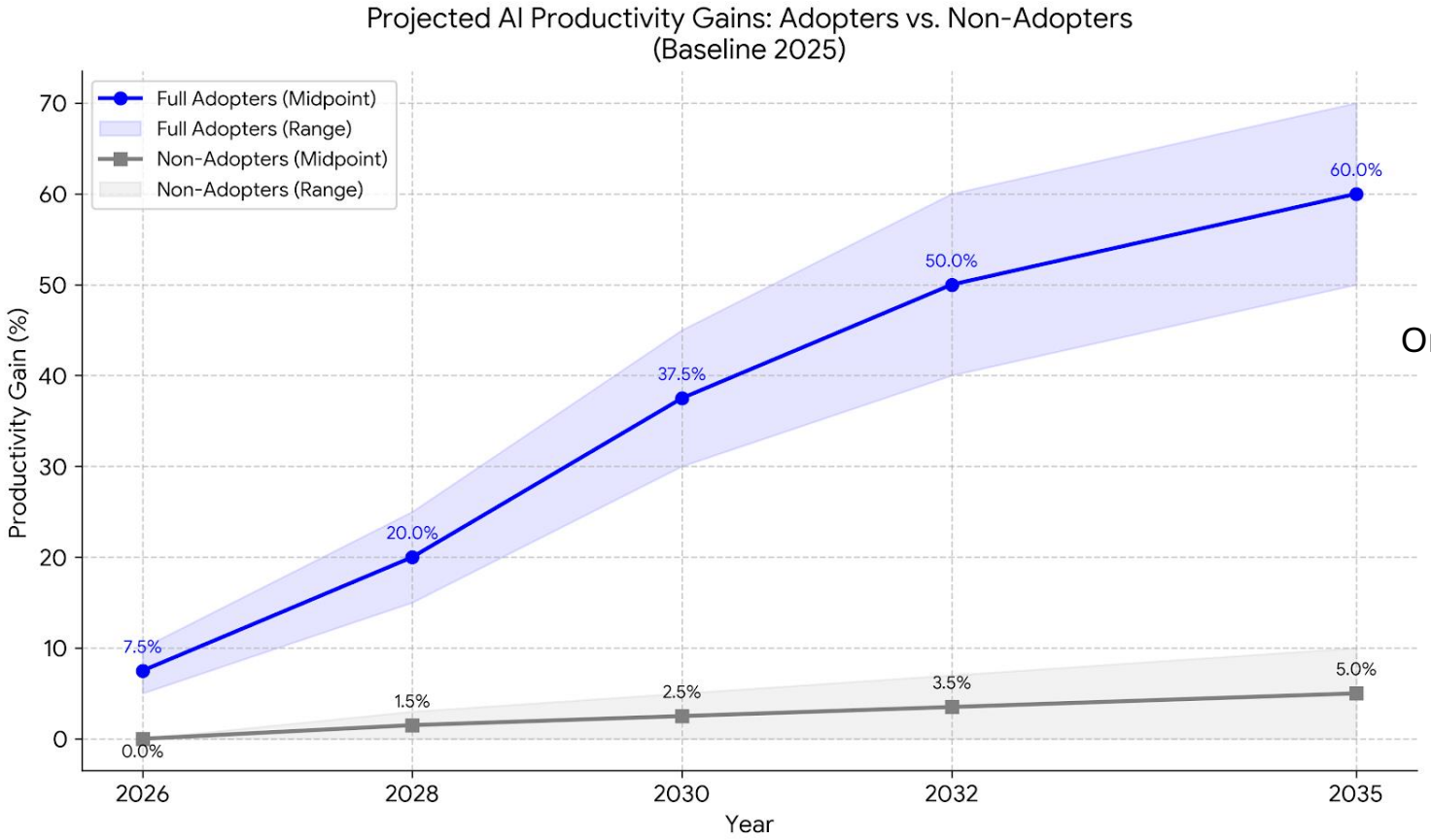
In 3-5 years

Category	Threat	Probability	Severity	Risk Level
External	Autonomous multi-agent attacks (adaptive, coordinated)	High	Critical	Critical
External	Deepfake and synthetic media fraud at global scale	High	Major–Critical	Critical
Internal/ External	AI-driven physical attacks (robots, IoT, VR systems)	Medium–High	Critical	Critical
Internal	Shadow models & unmonitored internal AI deployments	High	Major	High
Internal/ External	Supply-chain/model compromise (poisoned OSS models)	Medium	Major	High
Internal	Compliance failure under new AI regulations	Medium–High	Major	High
External	Autonomous multi-agent attacks (adaptive, coordinated)	High	Critical	Critical

Building blocks for a secure AI future

Category	Core Strategy	Goal	Key Actions
Governance	Clear ownership & rules	Define responsibility and boundaries	• Assign AI risk owner • Approve AI use policy • Keep risk register
Secure Design	Build security into architecture	Prevent misuse and leaks by design	• Use least-privilege access • Isolate data/models • Apply secure-by-design principles
Data Protection	Safeguard sensitive info	Avoid data loss, breaches, and legal issues	• Classify & encrypt data • Redact before prompts • Enable DLP controls
People & Culture	Train and guide safe use	Turn users into a defense layer	• Educate on AI risks • Ban shadow AI • Provide quick “Do’s & Don’ts”
Vendors & Tools	Secure your ecosystem	Ensure trusted, compliant third-party AI	• Vet vendors • Use secure gateways • Add AI clauses to contracts
Monitoring & Testing	Detect issues early	Keep systems safe as AI evolves	• Log usage • Red-team regularly • Have incident playbooks
Future Readiness	Stay ahead of regulation & physical AI	Prepare for next-gen risks (robots, IoT, VR)	• Track new laws • Audit regularly • Extend security to devices

And what if I decide not to use AI?



Or maybe..

First steps towards being safe

1. Define Ownership

Assign an AI Risk Owner (CISO, CIO) and a governance team to establish clear accountability from the start.

2. Set Basic Rules

Approve an internal AI Use Policy defining what's allowed and which tools are approved to stop unsafe use.

3. Protect Your Data

Identify sensitive data and decide what must **never** enter AI tools to prevent leaks and compliance issues.

4. Pick Trusted Tools

Use secure, enterprise-grade AI vendors with clear data terms to reduce vendor and privacy risk.

5. Educate & Monitor

Train staff on safe AI use and review logs regularly to build awareness and detect misuse early.

Ready for AI ? Our booklets

We have created short AI guides for decision-makers – and new ones to come!

Page: www.ai4leaders.ch/ -> Downloads
Password: **Booklets25**

(for now, only 2 available on the EN page)

..and let's connect on