



AI and Cybersecurity

How organizations can adopt AI
with greater confidence

Dr. Dimitrios Patsos

CISSP, CISM, CDPSE, CCSK, AAISM,

Global Black Belt, Security & Frontier AI

Microsoft Corporation



What Microsoft observes at global scale

>165tr

security signals
daily

15k+

partners
in our security ecosystem

35,000

full-time equivalent
security engineers
employed worldwide

31m

identity risk detections
analyzed in an average day

5.2bn

emails screened daily
on average to protect users
from malware and phishing

~4.7m

net new malware
file blocks
every day

- Personalize phishing and generate synthetic identity





The shift: same goals, faster operating model

Attackers still target identity, credentials, and access — AI changes the operating tempo

“What has changed is not the goal of the attacker, but the scale and speed at which they operate.”



Scale of access

AI-powered tooling amplifies attacks against the identity plane.

Amplify



Speed of execution

Automation and orchestration compress the time from compromise to impact.

Automate



Expanded identity surface

AI-driven services and machine-to-machine access create more paths to abuse.

Expand surface

- Permissions didn't change. Reachability did.

BEFORE

Employee accesses data manually.



TODAY

AI accesses data on behalf of the employee.

RISK

AI can discover, correlate and summarise information humans never would.

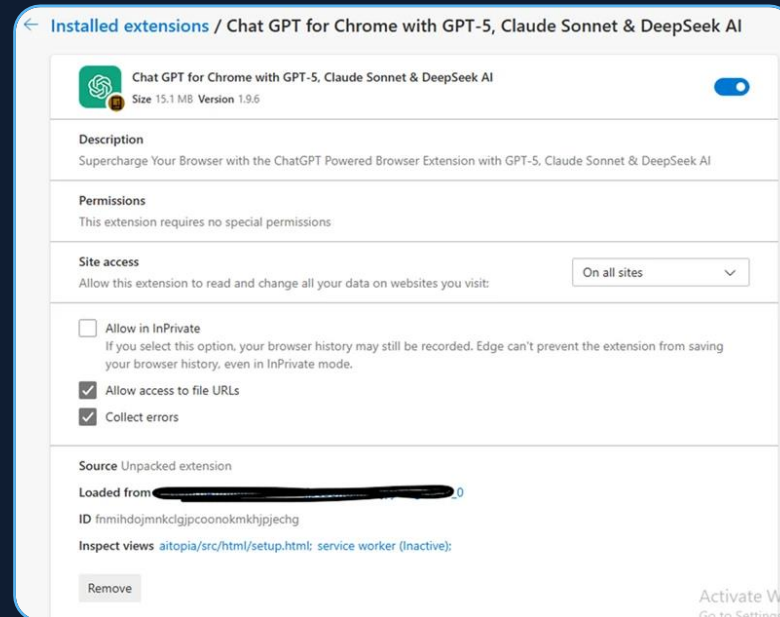
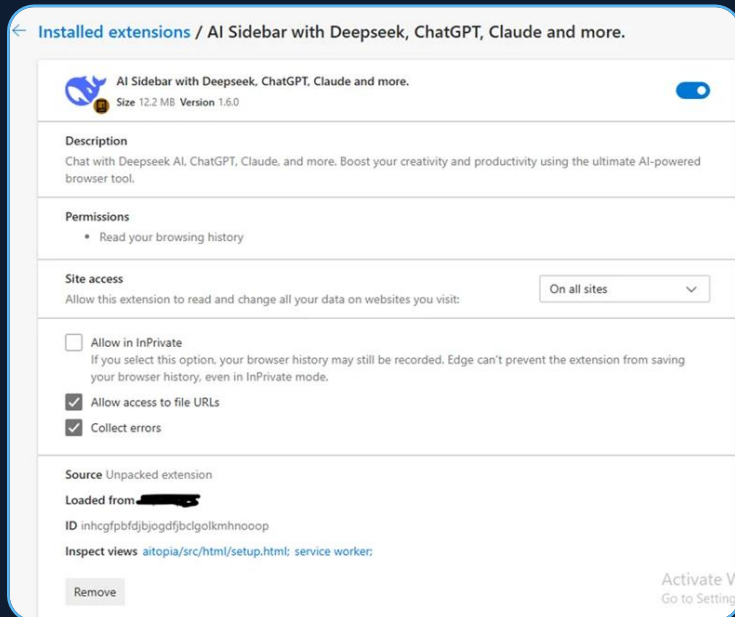
When a trusted-looking AI tool becomes the risk

Install

Grant permissions

Extension operates normally

Data collected & exfiltrated



KEY OBSERVATIONS

- SEO manipulation + bot reviews pushed extension to #1
- Earned "Featured" badge,
- Consent disguised as "anonymous analytics"
- **900,000 installs**
- **20,000 enterprise tenants**



The agentic defense stack: Five risk classes that define the 2026 attack surface

Effective protection requires consistent enforcement of controls across deployed agents. Malicious or policy violating use includes attempts to generate harmful or prohibited content. Illustrative examples include drafting phishing emails or generating fraudulent invoices.

Risk area	What attackers do	What defenders need
Prompt and intent manipulation	Inject instructions through prompts, content, files, or memory to redirect the agent	Prompt injection detection, payload inspection, intent validation, output review
Sensitive data exposure	Drive the agent to read or return content beyond its sanctioned scope	Sensitivity-aware retrieval, response inspection, data loss prevention (DLP) on inputs and outputs, memory scoping
Identity and privilege compromise	Impersonate a sub-agent, reuse stolen credentials, or chain privileges across agents	Verifiable agent identity, mutual authentication, scoped credentials, least privilege
Excessive agency	Coax the agent into chaining tools or actions beyond sanctioned behavior	Tool allow-listing, runtime gating, action policy, anomaly detection
Operational integrity	Tamper with configuration, system prompts, memory, training data, supply chain, or logs	Immutable logs, signed configuration, change governance, supply chain attestation



Every agent needs an identity, a purpose and an owner

Inside the agent attack surface:

- Malicious link injection routed through model and tool outputs (52%)
- Jailbreak and safety bypass attempts (16%)
- Tool and agent abuse (12%)
- Identity and credential abuse (10%)

Agents create different challenges from traditional software

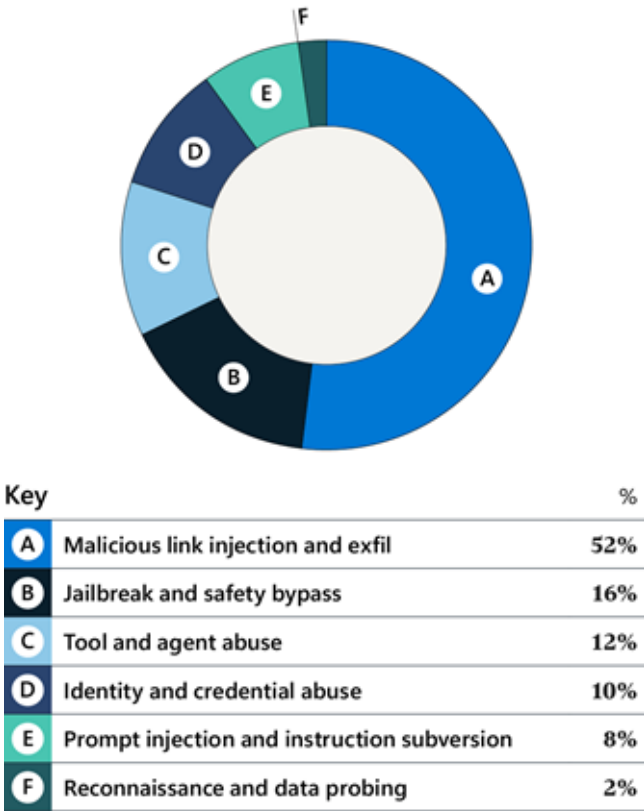
1. Ungoverned agent proliferation
2. Overprivileged access
3. No human accountability

A risk-based approach to agent governance

1. Start with identity. Every agent must have a distinct, auditable identity that carries metadata about who created the agent, its purpose, and its sponsor.
2. Enforce human accountability. When an agent’s sponsor changes, agent accountability must be re-evaluated and transferred automatically to a new sponsor. No environment should ever have an agent without a human to answer for it.

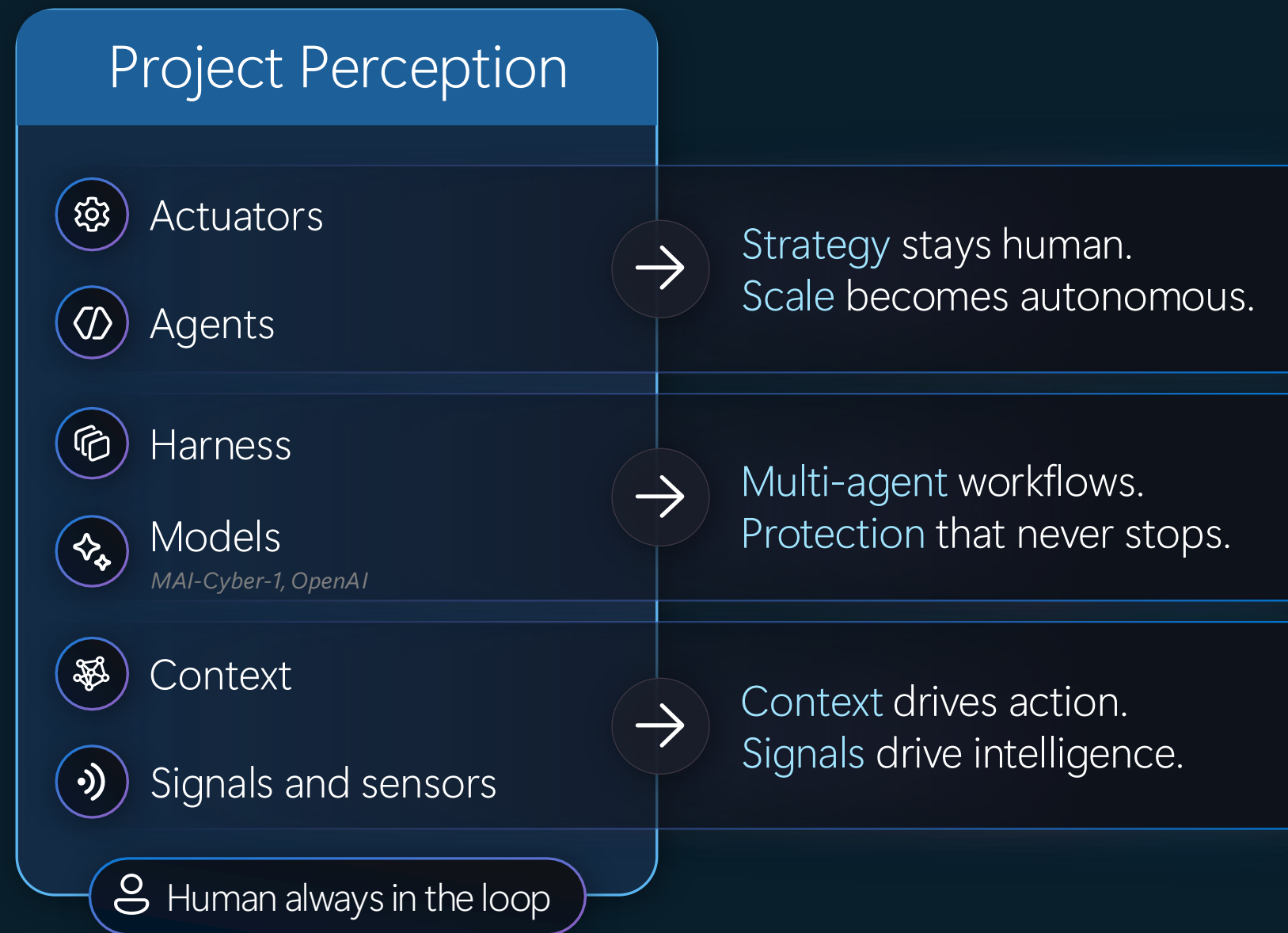
Figure 2. Top attack vector categories in telemetry data on AI workloads

Malicious link injection, tool and AI agent section abuse, and prompt subversion account for the majority of observed attack activity on AI workloads.



Source: Telemetry data on AI workloads (90 day window, 2026).

AI can also strengthen the defender





Five practical steps for secure AI adoption

1

Know which AI tools and agents are being used

2

Assign Ownership and Accountability

3

Protect identities and apply least privilege

4

Classify and protect sensitive information

5

Use AI to improve detection and response, with human oversight



Thank you!