Security of Agentic AI Systems

Tip of the Iceberg

What every business leader needs to know about the new era of AI-driven cyber threats

April 2026

THE STAKES

Security is now the #1 blocker to scaling agentic AI.

62%

of organizations cite security and risk as the primary barrier to scaling agentic AI.

38%	cite technical limitations
38%	cite regulatory uncertainty
32%	cite responsible-AI tooling gaps

The technology is ready. The defenses are not.

Stanford AI Index 2026

What Changed — Chatbots Said Things. Agents Do Things.

Generative AI (2022–2024)

AI Assistants — built for user interaction

- Reactive — answers prompts
- Stateless — no memory of prior turns
- Read-only — outputs text
- Limited autonomy
- Worst case: embarrassing or wrong answer

Agentic AI (2025 →)

Built for autonomous collaboration

- Autonomous — pursues goals on its own
- Persistent memory across sessions
- Calls tools, APIs, executes code, sends email
- Dynamic, multi-agent orchestration
- Worst case: real-world consequences at machine speed

The blast radius of an AI failure just got physical.

This Is Not Theoretical — Real Damage in Production

EchoLeak

Microsoft 365 Copilot · June 2025

Zero-click data exfiltration via a single email. No user action required.

Asana AI

Tenant isolation flaw · 2025

Cross-tenant data contamination affected up to 1,000 enterprises.

Devin AI

Coding agent · Late 2025 · $500 to break

Defenseless against prompt injection. Could leak tokens, install C2 malware.

GitHub Copilot

RCE via prompt injection · 2025

Remote code execution on developer machines — tens of millions of potential targets.

Sources: Aim Labs, NIST NVD, CoSAI, Rehberger 2025, Adversa Top AI Incidents 2025

THE NEW ATTACK SURFACE

Five surfaces every agent introduces.

Prompt Injection

Hostile instructions hidden in any text the agent reads like emails, web pages, docs, tool output.

Memory Poisoning

Persistent false beliefs planted in long-term memory. Sleeper agents that wake weeks later.

Tool Misuse

Legitimate tools weaponized into sending emails, deleting data, transferring funds.

Supply Chain

Malicious MCP servers, poisoned skills, compromised model packages.

Identity Abuse

Stolen agent credentials look like legitimate traffic. Fastest-growing attack vector.

Prompt injection appeared in 73% of production AI deployments in 2025. — OWASP GenAI Security Project


02
Frameworks &
Practical Defenses

FRAMEWORKS

Three lenses. Use all three.

OWASP

Agentic AI Top 10 (2026)

ANSWERS:

What can go wrong

MITRE ATLAS

Adversarial Threat Landscape

ANSWERS:

How adversaries attack

NIST AI RMF

AI 600-1 + Agentic Profile

ANSWERS:

How to govern the program

They're not competing — they're complementary. Map your controls against all three.

Framework 1 — OWASP Agentic AI Top 10 (2026)

ASI01 — Agent Goal Hijack

ASI02 — Memory Poisoning

ASI03 — Tool Misuse

ASI04 — Privilege Compromise

ASI05 — Indirect Prompt Injection

ASI06 — Supply Chain & Skills

ASI07 — Identity & Auth Abuse

ASI08 — Insecure Inter-Agent Comms

ASI09 — Excessive Autonomy

ASI10 — Rogue Agents

First peer-reviewed framework for autonomous AI

Released Dec 2025 · 100+ security experts
Endorsed by NIST, Microsoft, NVIDIA

Top 3 enterprises hit hardest: Memory Poisoning, Tool Misuse, Privilege Compromise.

Framework 2 — MITRE ATLAS

16 tactics

84 techniques

42 real-world case studies

Maps adversary behavior the same way ATT&CK does for IT — but for AI.

- Feb 2026 update added agent-specific techniques: Publish Poisoned Tool, Escape to Host, Modify Agentic Configuration.
- 70% of ATLAS mitigations map to existing security controls — your SOC can integrate today.
- Use it for red teaming, threat modeling, and tabletop exercises.

atlas.mitre.org · Free tools: ATLAS Navigator, Arsenal

Framework 3 — NIST AI RMF

GOVERN

Culture, accountability, policy. Who owns AI risk?

MAP

Identify context, risks, intended outcomes per system.

MEASURE

Quantify risks. Test, evaluate, red-team continuously.

MANAGE

Prioritize, respond, remediate. Treat AI risks like any other.

Agentic Profile coming Q4 2026 (NIST CAISI). Until then: pair AI 600-1 with the OWASP Agentic Top 10.

Containment is non-negotiable.

The principle

Assume the agent will do the wrong thing. Build the room so wrong is contained.

Least privilege

Restricted file system, network, OS calls

Resource and rate limits — block runaway loops

Containers alone are not enough

TEEs

Hardware-isolated, remote-attestable. Highest assurance.

MicroVMs

gVisor, Kata Containers. Stronger than containers.

OS Sandboxes

seccomp, SELinux. Restrict syscalls at kernel level.

Containers

Necessary, not sufficient. Easily escaped if misconfigured.

DEFENSE IN DEPTH

No single layer holds. All of them together do.

Governance	Policy, approval gates, audit, accountability. Zero-trust architecture. Clear ownership + kill switches.
Identity & Auth	Per-agent identity, scoped tokens, OAuth + token exchange. Agents are first-class identities with least privilege.
Sandboxing	TEEs, microVMs, OS sandboxes, resource limits. Execution sandboxes for model runtime.
I/O Hygiene	Schema validation, prompt-injection detection, output sanitization. Data minimization by design.
Observability	Centralized logs, correlation IDs, OpenTelemetry, prompt-to-action tracing. Continuous monitoring.

Each layer fails independently. Stacked, they buy you time and forensics.

CALL TO ACTION

What to do Monday.

01 **Inventory your agents**

Find every agent, every tool it can call, every credential it holds.

02 **Pick a framework and map controls**

OWASP Agentic Top 10 is the fastest start. ATLAS for red team. NIST for governance.

03 **Sandbox by default**

No agent runs in production without resource limits, scoped tokens, and proper isolation.

04 **Approve MCP servers like software**

Code signing, SBOMs, allowlists. No anonymous installs. Treat marketplace skills as untrusted.

05 **Red team deployed agents**

Prompt injection success rates exceed 85%. Find your gaps before someone else does.

06 **Trace prompt → tool → action**

If your SIEM can't link a user prompt to a downstream API call, incident response is blind.

03

Mythos Moment&
What This Means

What Is Claude Mythos?

A structural shift in the cyber threat landscape

A new kind of AI

Anthropic revealed that Claude Mythos can autonomously find and exploit security vulnerabilities — not just by finding critical gaps, but by chaining low-severity vulnerabilities across browser, OS, network and application layers into working attacks. No security expertise required.

It wasn't designed to do this

Mythos was built as a general coding tool. The offensive capability emerged on its own. Any sufficiently advanced AI may develop similar capabilities.

Defenders got first access — for now

Anthropic gave restricted access to major partners through Project Glasswing. This head start is expected to last ~12 months.

83% first-attempt exploit success rate

181 working attacks vs. 2 before Mythos

27 yrs oldest vulnerability discovered

~$0 cost to launch a sophisticated attack

The same AI that can protect your business can also be turned against it. Defenders who act now have a window measured in months.

Why This Is Different

Speed

BEFORE

Skilled attackers needed weeks to develop a working exploit.

NOW

Mythos produces working attacks in hours — faster than any organisation can patch.

Your 30-day patch cycle is now dangerously slow.

Impact

BEFORE

Risk was assessed vulnerability-by-vulnerability. A minor flaw was a minor risk.

NOW

Mythos chains multiple small weaknesses into a single devastating attack.

Your current risk scoring model needs to be rebuilt.

Scale

BEFORE

Launching a sophisticated attack required rare, expensive expertise.

NOW

Comparable tools will enable anyone with basic skills to execute sophisticated attacks.

The threat landscape is about to expand dramatically.

Mythos Timeline Projection (Now – 12 Months)

HORIZON 1 · Weeks 1–4

CVE Flood

Vulnerability volume will overwhelm patch capacity. Vulnerabilities that took weeks to weaponize can now be exploited in hours.

The question to ask:

"Do you have a clear plan to deal with the increased CVEs?"

HORIZON 2 · Months 1–3

The Mythos Moment

Autonomous zero-day discovery changes the game. Frontier AI models are removing the barrier to sophisticated cyberattacks.

The question to ask:

"How far are you with adopting Zero-Trust architecture?"

HORIZON 3 · Months 3–12

Baseline Increase

Security posture must permanently change to be proactive. Resilient zero-trust architectures become the new standard.

The question to ask:

"Is your security operations automated and AI-enabled?"

Four Consequences Everyone Must Understand

1

Security budget assumptions are out of date

Defenses built for human-speed attacks are not designed for AI-speed threats. The cost of a breach and investment needed will fundamentally change.

2

Risks we thought were low may now be critical

AI can combine several minor weaknesses into a catastrophic attack. Issues rated 'low priority' may now be the entry point.

3

Tool proliferation within 12–18 months

Today only the most sophisticated actors have access. Within 18 months, comparable tools will be widely available to criminal groups.

4

Doing nothing is the highest-risk strategy

Regulations, cyber insurers and stakeholders will demand action. Operational disruption resilience becomes a competitive differentiator.

What can you do?

01

CTEM

Continuous Threat Exposure Management

- Scope, discover, prioritize, validate, mobilize — continuously
- Continuous exposure reduction, not periodic scans
- Unify ASM, BAS, and pentesting under one rhythm

3x
less likely to be breached (Gartner)

02

Graph Analytics

Attack-path-based prioritization

- Map identity, network, and CVE relationships in one graph
- Find toxic combinations and choke points at scale
- XM Cyber, Tenable One, BloodHound, MSEM

98%
noise cut via validated attack paths

03

Identity-first Zero Trust

Humans, machines, AI agents

- One IAM plane for every human, machine, and AI agent
- Workload identity; eliminate static secrets
- Least privilege per task; human gates on high-impact actions

82:1
AI-agent to human ratio (Palo Alto)

04

Defend at machine speed

Autonomous SOC and response

- AI-driven triage, investigation, and containment
- Closed-loop response across EDR and identity platforms
- Continuous AI red-teaming over static playbooks

22s
defender response window (Mandiant 2026)

05

Resilience by design

Assume breach; recover fast

- Immutable, air-gapped backups tested by recovery drills
- Quarterly ransomware tabletops with rollback plans
- RTO/RPO mapped to business processes; vendor-inclusive drills

96%
of ransomware hits target backups (Veeam)

Sources: Gartner CTEM 2024 | UK AI Safety Institute Mythos Preview evaluation | Mandiant M-Trends 2026 | Palo Alto Networks 2026 AI predictions | Veeam Ransomware Report

HOW WE CAN HELP

A

Executive Briefing

Briefing session for client board or CISO. Threat landscape mapped to client-specific industry. Board-ready framing.

Within 2 weeks

B

Cyber.AI & Agent Shield Demo

Live demonstration of AI-driven threat detection. Agent Shield governance. Integration roadmap.

Available now

C

Map → Assess → Act Programme

Full exposure mapping. AI threat assessment and architecture review. Remediation roadmap.

Kick off within 30 days

Agentic AI is the largest expansion of the enterprise attack surface in a decade.
Treat it that way.

Interested in exploring how these solutions can support your business?

Connect with IT One team to learn more.

IT1.BSA@accenture.com

SCG Chemicals

K. Piyachat Dhanaraks
piyachat.dhanaraks@itone.co.th

SCG Packaging

K. Kitti Boonyakitnotai
kitti.boonyakitnotai@itone.co.th

SCG CBM

K. Sunion Panichattra
sunion.panichattra@itone.co.th

SCG Corporate

K. Soraphot Markathub
soraphot.markathub@accenture.com