



CYBEXIS

WHITE PAPER | AI RISK

Pragmatic AI Risk Management: Identifying and Mitigating AI Threats

A proportionate methodology for securing the AI you use and defending against the AI used against you

October 2026

Clarity in complexity. Confidence in every decision.

At a glance

- AI threats now come from two directions: attackers using AI to scale intrusion and fraud, and attackers targeting the AI systems organisations deploy.
- Indirect prompt injection moved from research to live exploitation in 2026. Agents that combine private data, untrusted content and the ability to act are the highest-risk pattern.
- Existing risk methods work if adapted: inventory, tier by autonomy and impact, threat model with OWASP and MITRE ATLAS, then apply layered controls and continuous testing.
- Regulation is phasing in. EU AI Act transparency and general-purpose AI duties applied from August 2026, while high-risk obligations were deferred to December 2027 and August 2028.

Who this is for: CISOs, risk and governance leads, architects and product owners deploying generative AI, assistants or autonomous agents.

Information correct as of October 2026. Free to share with attribution to Cybexis.

1. The AI threat landscape in 2026

Two years ago most AI security discussion was speculative. That is no longer the case. Organisations now face AI risk on two fronts, and a pragmatic approach has to address both.

Attackers using AI

- **Automated intrusion.** In November 2025 Anthropic reported disrupting a campaign it attributed to a Chinese state-sponsored group, in which its coding agent was manipulated into performing an estimated 80 to 90% of tactical operations against around 30 organisations, with humans intervening at a handful of decision points [1]. Some researchers questioned the figures, but the direction of travel is clear.
- **Exploitation at scale.** The NCSC has warned that by 2028 AI-enabled capabilities will likely be used to exploit known vulnerabilities in legacy technology at scale across critical infrastructure [2]. The window between disclosure and exploitation will keep shrinking.
- **Convincing social engineering.** Deepfake video and voice have been used to authorise fraudulent payments, most notably the 2024 case in which an engineering firm's Hong Kong office transferred around US\$25 million after a video call with a synthetic "CFO". AI-written phishing is now fluent, personalised and cheap.

Attackers targeting AI

- **Indirect prompt injection in the wild.** In March 2026 Palo Alto Networks' Unit 42 reported live web pages carrying hidden instructions aimed at AI agents, including attempts to trigger payments and leak system prompts [3]. The Cloud Security Alliance concluded in April 2026 that indirect prompt injection had crossed from proof-of-concept to live exploitation [4].
- **From bad output to code execution.** Vulnerabilities disclosed in 2026 showed prompt injection in agent frameworks escalating to remote code execution, and crafted issue and pull request text hijacking coding agents running in CI pipelines [5].
- **Poisoned tools and memory.** Malicious "skills" and plug-ins for agent frameworks have been distributed through public marketplaces, and researchers have demonstrated attacks that write false facts into an agent's long-term memory for later use [5]. OWASP now treats memory and context poisoning as a distinct risk.
- **Data leakage and shadow AI.** The most common day-to-day harm remains staff pasting sensitive data into unapproved tools, or assistants with broad access surfacing information users should not see.

The "lethal trifecta"

A useful heuristic popularised by Simon Willison: an AI system is most exposed when it combines **access to private data**, **exposure to untrusted content** and **the ability to communicate or act externally**. Any one can be managed; all three together mean a single injected instruction can exfiltrate data or take unwanted action. Breaking at least one leg is the most effective design control available today.

2. The governance and standards landscape

There is no single AI security standard, but the main references are now mature enough to build on. Most organisations should anchor to one management framework and use the technical taxonomies for threat modelling and testing.

Reference	Type	Use it for
ISO/IEC 42001:2023	Management system	A certifiable AI management system that integrates with ISO/IEC 27001.
NIST AI RMF 1.0 and Generative AI Profile	Risk framework	Govern, Map, Measure, Manage functions; GenAI-specific risks and actions.

Reference	Type	Use it for
UK AI Cyber Security Code of Practice (DSIT, 2025)	Baseline principles	Lifecycle security expectations for developers and operators of AI systems.
OWASP Top 10 for LLM Applications (2026) and for Agentic Applications (2026)	Threat taxonomy	Application-level risks such as prompt injection, data leakage, goal hijack, tool misuse and rogue agents [6].
MITRE ATLAS	Adversary techniques	ATT&CK-style tactics and techniques against AI systems, for threat modelling and red teaming.
NCSC CAF v4.0	Regulatory outcome framework	Requires operators to consider new and emerging technologies, including AI, in risk assessment.
EU AI Act (as amended by the Digital Omnibus)	Regulation	Transparency and GPAI duties from 2 August 2026; high-risk duties from 2 December 2027 (Annex III) and 2 August 2028 (Annex I) [7].

Table 1. Key AI security and governance references.

For UK organisations without EU exposure, the AI Act still matters indirectly through customers and suppliers, and the deferral of high-risk obligations is a window to prepare rather than a pause. The classification rules did not change.

3. A pragmatic five-step methodology

The aim is proportionality: concentrate effort on the AI uses that can cause real harm, and let low-risk uses proceed quickly with light-touch guardrails. The method below extends familiar risk management practice rather than replacing it.

Step 1: Inventory every AI use

Record each AI capability in use, including embedded features in SaaS products, developer tools, internally built applications and agents, and staff use of public tools. For each, capture the owner, business purpose, model and provider, data accessed, tools and permissions granted, and who or what consumes the output. Discovery through network, identity and SaaS management logs usually finds far more than a staff survey.

Step 2: Tier by autonomy and impact

Tier	Typical pattern	Examples	Expected controls
1. Assist	Generates content a human reviews; no sensitive data	Drafting, summarising public material	Approved tools, acceptable use policy, awareness
2. Inform	Uses internal data to inform human decisions	Enterprise search, RAG assistants	Data permissions honoured, logging, output review, DLP
3. Act with approval	Proposes actions a human approves	Ticket triage, draft payments, code changes with review	Threat model, least-privilege tools, human-in-the-loop, red team testing
4. Act autonomously	Takes actions without per-action approval	Agents with email, finance, infrastructure or OT access	Full assessment, sandboxing, scoped credentials, kill switch, continuous monitoring, board visibility

Table 2. A simple AI risk tiering model. Raise the tier by one if the system processes special category data or supports a regulated or safety-related decision.

Step 3: Threat model the system, not the model

For Tier 3 and 4 uses, run a structured threat model. Draw the data flows from user and external content through the model, retrieval stores, memory and tools to the systems it can affect. Then walk each trust boundary using the OWASP LLM and Agentic Top 10 as prompts, and MITRE ATLAS for adversary techniques. Four questions catch most issues:

1. What untrusted content can reach the model, directly or indirectly (web pages, emails, documents, tickets, tool outputs)?
2. What can the system do with the authority it holds, and is that more than the task needs?
3. What would happen if the model followed an attacker's instruction instead of the user's?
4. How would we know, and how quickly could we stop it?

Step 4: Apply layered controls

Threat	Primary controls
Prompt injection and goal hijack	Treat all model input as untrusted; separate instructions from data; constrain tools so a hijacked agent cannot do serious harm; require approval for high-impact actions; filter and monitor outputs.
Excessive agency and tool misuse	Least-privilege, task-scoped credentials per agent; allow-listed tools and destinations; rate limits; no standing access to production or payments.
Data leakage	Enforce source-system permissions in retrieval; classify and exclude sensitive stores; DLP on prompts and outputs; contractual no-training terms with providers.
Supply chain (models, plug-ins, skills)	Approved sources only; review and pin versions; scan packages; inventory in the software bill of materials; vendor security assessment.
Memory and context poisoning	Restrict what can be written to persistent memory; provenance on stored facts; periodic review and reset; isolate memory between users.
AI-enabled social engineering	Out-of-band verification for payments and credential resets; phishing-resistant MFA; staff briefings on deepfakes; code words for urgent executive requests.
AI-accelerated exploitation	Faster patching of internet-facing systems; attack surface management; replace or isolate legacy technology; assume shorter time to exploit.

Table 3. Mapping AI threats to proportionate controls.

Step 5: Test, monitor and govern continuously

AI systems change behaviour when models, prompts, data or tools change, so point-in-time assurance decays quickly. Build adversarial test cases into release pipelines, red team Tier 3 and 4 systems before go-live and after material change, log prompts, tool calls and outputs with appropriate privacy controls, and alert on anomalous behaviour such as unusual tool sequences or data volumes. Report AI risk through the existing risk register with named owners, rather than in a separate silo.

4. Key challenges

- **There is no complete fix for prompt injection.** Research in 2026 showed adaptive attacks bypassing most published defences [5]. Controls must assume injection will sometimes succeed and limit the consequences.
- **Visibility.** AI features arrive inside existing SaaS products by default, often without a procurement event to trigger review.
- **Accountability for agent actions.** Identity and access models built for humans and service accounts do not cleanly describe an agent acting on behalf of a user with delegated authority.
- **Speed versus assurance.** Business pressure to adopt AI quickly can turn security into a perceived blocker. Tiering is the main tool for saying yes safely to most uses.
- **Skills.** Few teams yet combine security testing, data science and product knowledge. Upskilling existing threat modellers is usually faster than hiring.

5. Priorities for the next 12 months

1. **Publish a short AI acceptable use standard** and an approved tools list, so staff have a safe default.
2. **Complete an AI inventory and tier every use.** Revisit it quarterly.



3. **Threat model and red team every Tier 3 and Tier 4 system** before it goes live, focusing on the lethal trifecta.
4. **Harden identity for agents.** Unique, scoped, short-lived credentials; no shared or over-privileged service accounts.
5. **Update payment and access procedures for deepfakes.** Mandatory out-of-band verification for any urgent request to move money or change credentials.
6. **Accelerate fundamentals that AI-enabled attackers exploit:** patch latency on internet-facing systems, legacy exposure and phishing-resistant MFA.
7. **Choose a governance anchor** (ISO/IEC 42001 or NIST AI RMF) and integrate it with your existing ISMS and risk register.

AI does not require a new discipline of risk management so much as a disciplined application of the one organisations already have, adjusted for systems that can be persuaded by the content they read. Organisations that inventory honestly, focus on autonomy and impact, and design for failure will be able to adopt AI confidently while staying ahead of attackers who are doing the same.

About Cybexis

Cybexis is an independent UK cyber security advisory, assurance and risk management practice based in Reading. Its insights series shares practitioner perspectives on the frameworks, regulation and threats facing regulated and high-risk organisations. www.cybexis.co.uk | hello@cybexis.co.uk

This paper is for general information only and does not constitute legal or regulatory advice. Requirements, deadlines and threat activity change; always confirm current obligations with your regulator, contracting authority or certification body.

Sources and further reading

- [1] Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign (Nov 2025). <https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf>
- [2] Computer Weekly, Hostile states launched 200 attacks on UK infrastructure in five months, says NCSC chief (2026). <https://www.computerweekly.com/news/366644872/Hostile-states-launched-200-attacks-on-UK-infrastructure-in-five-months-says-NCSC-chief>
- [3] Vibe Graveyard, Unit 42 finds AI agents obeying poisoned web pages (summary of Unit 42 research, March 2026). <https://vibegraveyard.ai/story/unit42-web-indirect-prompt-injection-in-the-wild/>
- [4] Cloud Security Alliance, Indirect prompt injection goes operational (April 2026). <https://labs.cloudsecurityalliance.org/research/csa-research-note-indirect-prompt-injection-in-the-wild-2026/>
- [5] The AI agent security problem stopped being hypothetical in 2026 (Level Up Coding). <https://levelup.gitconnected.com/the-ai-agent-security-problem-stopped-being-hypothetical-in-2026-c0b81c438d72>
- [6] OWASP GenAI Security Project, Top 10 for Agentic Applications for 2026. <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026>
- [7] Cloud Security Alliance, EU AI Act high-risk deadline: deferred, not cancelled (2026). <https://labs.cloudsecurityalliance.org/research/csa-research-note-eu-ai-act-high-risk-deadline-omnibus-20260/>
- [8] Process Excellence Network, EU AI Act 2026: what changed and what deployers must still do. <https://www.processexcellencenetwork.com/ai/articles/eu-ai-act-2026-what-actually-changed-and-what-deployers-still-must-do>