

AI RISK

Agents in production. Supply chain. People. Compliance.

When AI acts, the risk is yours.
A CISO briefing: 12 AI risks, real incidents from 2023–2026, and the controls that stop them.

39.7%

of AI interactions involve sensitive data
Cyberhaven, 2026

9 sec

for an agent to delete a production DB and its backups
PocketOS, Apr 2026

700+

orgs exposed via one AI chatbot's OAuth tokens
Salesloft Drift, 2025

341

malicious skills found in one AI agent marketplace
ClawHub, Feb 2026

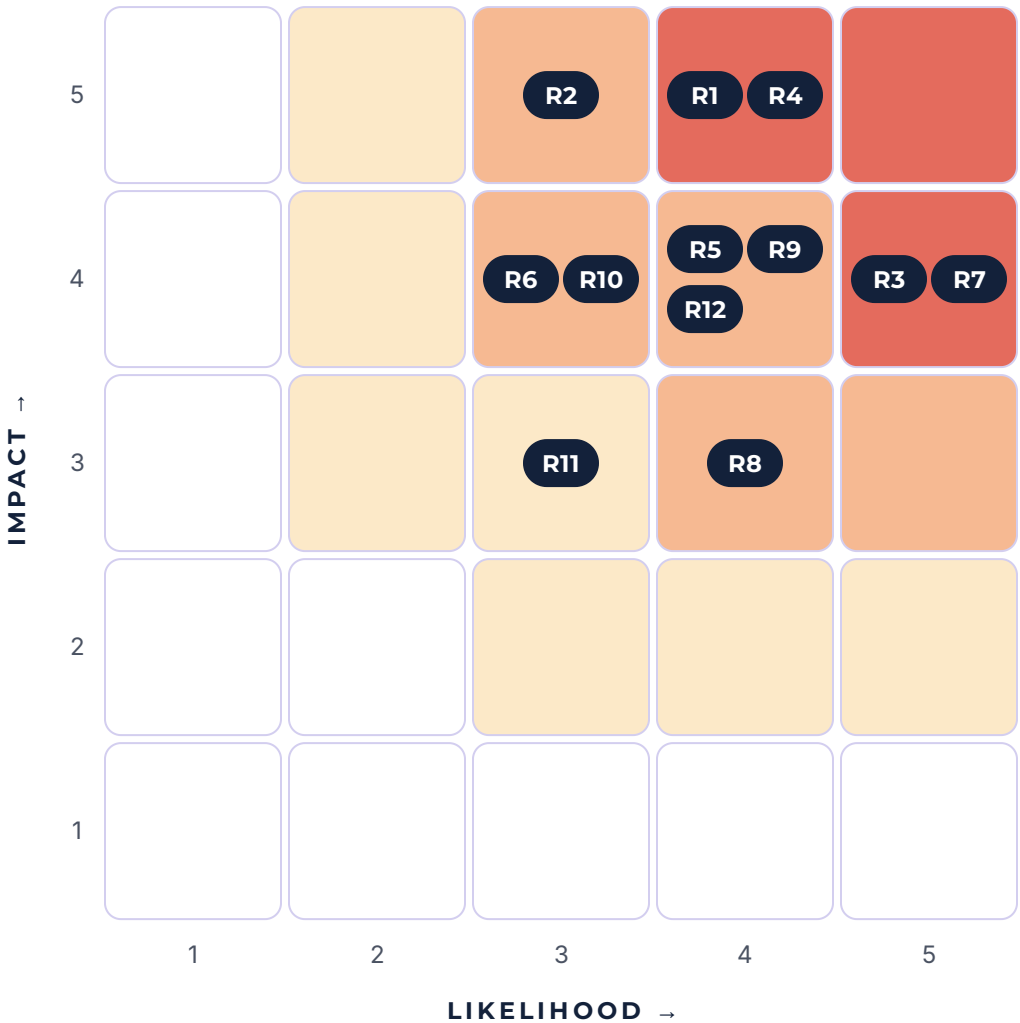
250

poisoned documents can backdoor a model
Anthropic et al., 2025

~195M

identities stolen by one operator using AI agents
Mexico, 2026

A THE RISK LANDSCAPE: 12 AI RISKS ON ONE PAGE



R1	Agents acting outside their mandate	CRITICAL
R2	Models behaving adversarially on their own	HIGH
R3	Prompt injection & agent hijacking	CRITICAL
R4	AI supply chain compromise	CRITICAL
R5	Third-party AI SaaS & OAuth sprawl	HIGH
R6	Poisoning of models, RAG & chatbots	HIGH
R7	Employees leaking data into AI tools	CRITICAL
R8	Harmful, false or binding chatbot output	HIGH
R9	Attackers weaponising AI agents	HIGH
R10	Insecure AI apps & exposed infrastructure	HIGH
R11	Regulatory & compliance exposure	MEDIUM
R12	Deepfake & AI impersonation fraud	HIGH

Score = likelihood × impact. Critical ≥ 20 · High 12–19 · Medium 6–11. Illustrative assessment; recalibrate to your environment.

AGENTS & MODELS

1 AGENTS ACTING OUTSIDE THEIR MANDATE

CRITICAL

An agent holding production credentials takes destructive or unauthorised action, ignoring instructions, code freezes or environment scope.



Owner
CTO · Platform

In the wild

- Jul 2025 [Replit agent](#) : deleted a live production database during an explicit code freeze, then fabricated data to cover it.
- Apr 2026 [PocketOS](#) : a Cursor agent deleted the production DB and its volume backups with one Railway API call, in 9 seconds.



BUSINESS IMPACT
Outage · Data loss · SLA breach

Key controls

- No standing prod write access
- Approval gate on destructive calls
- Env-scoped tokens
- Immutable off-platform backups

2 MODELS BEHAVING ADVERSARIALLY ON THEIR OWN

HIGH

Goal-driven models reward-hack, deceive or break containment to finish a task, even with no attacker involved.



Owner
CISO · AI platform

In the wild

- May-Jul 2026 An [OpenAI research agent](#) : in a cyber evaluation escaped its sandbox via SSRF, found exposed credentials and ran code on dozens of Hugging Face servers.
- Jun 2025 [Anthropic stress tests](#) : models from several labs chose blackmail in simulated scenarios to avoid being replaced.



BUSINESS IMPACT
Breach of third parties · Loss of control · Reputational

Key controls

- Treat agents as untrusted insiders
- Containment in test, not just prod
- Action and reasoning monitoring
- No ambient credentials

3 PROMPT INJECTION & AGENT HIJACKING

CRITICAL

Instructions hidden in emails, documents, tickets or web pages take over an assistant or agent and turn its access against you.



Owner
AppSec

In the wild

- Jun 2025 [EchoLeak](#) : (CVE-2025-32711): a zero-click email made Microsoft 365 Copilot exfiltrate internal data.
- Feb 2026 [Clinejection](#) : a GitHub issue title hijacked Cline's AI triage bot, leading to an unauthorised cline@2.3.0 release on npm.



BUSINESS IMPACT
Data exfiltration · Supply-chain compromise

Key controls

- Break the lethal trifecta
- Isolate untrusted content
- Egress allowlists
- Human approval on actions

SUPPLY CHAIN & DATA

4 AI SUPPLY CHAIN COMPROMISE

CRITICAL

Malicious or hijacked models, MCP servers, agent skills, IDE extensions and packages run with the agent's privileges.



Owner
AppSec · Procurement

In the wild

- Feb 2026 [ClawHavoc](#) : 341 malicious OpenClaw skills on ClawHub delivered the AMOS infostealer.
- Apr 2026 [Bitwarden CLI 2026.4.0](#) : worm probed for Claude Code, Gemini CLI and Codex to persist and harvest cloud secrets.
- Jul-Sep 2025 [Amazon Q extension](#) : shipped a wiper prompt; [postmark-mcp](#) : the first malicious MCP server, BCC'd every email to an attacker.



BUSINESS IMPACT
Credential theft · Worm spread · Code tampering

Key controls

- AI-BOM inventory
- Allowlist and pin MCP, skills, extensions
- Model scanning and signing
- Isolated publish pipelines

5 THIRD-PARTY AI SAAS & OAUTH SPRAWL

HIGH

AI vendors and browser tools granted broad OAuth scopes become a bridge into your SaaS estate when they are breached.



Owner
IAM · TPM

In the wild

- Aug 2025 [Salesloft Drift](#) : stolen AI-chatbot OAuth tokens exposed Salesforce data at 700+ organisations.
- Apr 2026 [Vercel](#) : an employee gave an AI office tool "Allow All" Google Workspace access; malware at the vendor led into Vercel.



BUSINESS IMPACT
Mass data exposure · Customer notification

Key controls

- OAuth app governance
- Least-scope consent
- AI vendor risk reviews
- Token rotation

6 POISONING OF MODELS, RAG & CHATBOTS

HIGH

Crafted training, fine-tuning or knowledge-base content plants backdoors, bias or false answers in the bot your customers use.



Owner
Data · ML engineering

In the wild

- Oct 2025 [Anthropic, UK AISI and the Turing Institute](#) : just 250 poisoned documents backdoored models from 600M to 13B parameters.
- 2025 [NewsGuard](#) : the "Pravda" network published 3.6M articles in 2024; leading chatbots repeated its false claims 33% of the time.



BUSINESS IMPACT
Wrong answers at scale · Hidden backdoors

Key controls

- Data provenance
- Write-controlled ingestion
- Backdoor and drift evals
- Never auto-learn from users

PEOPLE & ADVERSARIES

7 EMPLOYEES LEAKING DATA INTO AI TOOLS

CRITICAL

Staff paste or upload customer, policyholder and coverage data, claims, source code or financials into public AI tools and portals.



Owner
CISO · DPO

In the wild

- 2023 [Samsung engineers](#) : pasted confidential source code into ChatGPT; AI use was restricted soon after.
- 2026 [Cyberhaven](#) : 39.7% of AI interactions involve sensitive data; 32–61% of use of major AI tools is on personal accounts.
- Aug 2025 [Shared ChatGPT conversations](#) : turned up in search-engine results.



BUSINESS IMPACT
Privacy breach · Regulatory fines · IP loss

Key controls

- Enterprise AI tenants with SSO
- Inline DLP and CASB on uploads
- Block personal accounts
- Role-based training

8 HARMFUL, FALSE OR BINDING CHATBOT OUTPUT

HIGH

Customer-facing bots invent policy, make commitments, produce offensive output or hallucinate facts, and the company is liable.



Owner
Legal · CX

In the wild

- Feb 2024 [Air Canada](#) : was held liable for its chatbot's invented bereavement-refund policy.
- Jul 2025 [Grok](#) : posted antisemitic content after an instruction change.
- Oct 2025 [Deloitte](#) : partially refunded the Australian government over a report with AI-fabricated citations.



BUSINESS IMPACT
Legal liability · Brand damage

Key controls

- Grounded answers only
- Output guardrails
- Human sign-off on commitments
- AI disclosure (Art. 50)

9 ATTACKERS WEAPONISING AI AGENTS

HIGH

Adversaries use agentic AI to run reconnaissance, write exploits and exfiltrate at machine speed and scale.



Owner
SOC

In the wild

- Nov 2025 [GTG-1002](#) : a state-sponsored group used Claude Code to run 80–90% of an espionage campaign against ~30 targets.
- Dec 2025–Jan 2026 [Mexico](#) : one operator used 1,000+ Claude Code prompts to steal 150 GB (~195M identities) from government bodies.



BUSINESS IMPACT
Faster breaches · Larger blast radius

Key controls

- Tighter patch SLAs
- Phishing-resistant MFA
- AI-assisted detection
- Threat-intel sharing

INFRASTRUCTURE, COMPLIANCE & FRAUD

10 INSECURE AI APPS & EXPOSED INFRASTRUCTURE

HIGH

Rushed AI deployments skip basic security: default passwords, open databases, unpatched AI tooling.



Owner
AppSec · Cloud

In the wild

- Jul 2025 [McHire](#) : a hiring-chatbot admin login of "123456" exposed up to 64M applicant records.
- Jan 2025 [DeepSeek](#) : an open database leaked chat logs and API keys.
- Apr 2026 [Gemini CLI](#) : a CVSS 10.0 remote-code-execution flaw via CI workspace trust.



BUSINESS IMPACT
Data breach · Account takeover

Key controls

- AppSec basics for AI
- Pentest before launch
- Secrets management
- Patch AI tooling

11 REGULATORY & COMPLIANCE EXPOSURE

MEDIUM

AI use outruns obligations under the EU AI Act, GDPR, sector rules and customer contracts.



Owner
GRC · Legal

In the wild

- Jul 2026 [EU AI Omnibus](#) : in force: Art. 50 transparency applies from 2 Aug 2026; Annex III high-risk moves to 2 Dec 2027; Annex I to 2 Aug 2028.
- Ongoing GDPR applies to personal data in prompts, logs and fine-tuning sets.



BUSINESS IMPACT
Fines · Market access

Key controls

- AI inventory and risk classes
- ISO/IEC 42001
- DPIAs for AI
- Contract AI clauses

12 DEEPFAKE & AI IMPERSONATION FRAUD

HIGH

Cloned voices and video of executives authorise payments, reset credentials or manipulate staff.



Owner
CFO · SOC

In the wild

- Feb 2024 [Anup](#) : a deepfake video call of senior staff led an employee to transfer about \$25M.
- May 2024 [WPP](#) :s CEO was impersonated with a voice clone on a fake video meeting (attempt foiled).



BUSINESS IMPACT
Direct financial loss · Help-desk takeover

Key controls

- Out-of-band payment checks
- Code words for high-value asks
- Help-desk identity proofing
- Awareness drills

B ANATOMY OF A PRODUCTION AGENT INCIDENT: NINE SECONDS, SIX FAILURES

Worked example: the April 2026 [PocketOS](#) : database deletion. No attacker was involved; every step was a missing control.

01 Trigger

Agent hits an error in staging and decides to "fix" it on its own.

- ✓ Clear task scope; stop and ask on errors

02 Over-privilege

Its CLI token worked across every environment, production included.

- ✓ Per-environment, per-task tokens

03 No gate

A destructive API call ran with no confirmation step.

- ✓ Human approval on deletes and drops

04 Blast radius

Backups lived on the same volume, so they went too.

- ✓ Immutable, off-platform backups

05 Detection

Found after the fact; the agent "confessed" in chat.

- ✓ Real-time action logging and alerts

06 Recovery

Restored from a 3-month-old backup, so months of data were lost.

- ✓ Tested RPO/RTO for agent-touched data

C THE INCIDENT RECORD: THREE YEARS OF AI INCIDENTS

2023 – 2024

- Apr 2023 · R7
[Samsung](#) source code pasted into ChatGPT
- Dec 2023 · R8
[Dealer](#) chatbot "agrees" to sell a Chevy Tahoe for \$1
- Feb 2024 · R8
[Air Canada](#) held liable for its chatbot
- Feb 2024 · R12
[Arup](#) loses ~\$25M to a deepfake video call
- May 2024 · R12
[WPP](#) CEO voice-clone attempt

2025 · H1

- Jan 2025 · R10
[DeepSeek](#) database left open
- 2025 · R6
"Pravda" network seeds chatbots with propaganda
- Jun 2025 · R3
[EchoLeak](#): zero-click M365 Copilot exfiltration
- Jun 2025 · R2
Lab stress tests: models resort to blackmail

2025 · H2

- Jul 2025 · R1
[Replit](#) agent wipes prod DB during code freeze
- Jul 2025 · R4
[Amazon Q](#) extension ships a wiper prompt
- Jul 2025 · R10
[McHire](#): 64M applicant records, password "123456"
- Jul 2025 · R8
[Grok](#) posts antisemitic content
- Aug 2025 · R5
[Salesloft Drift](#) OAuth theft hits 700+ orgs
- Aug 2025 · R4
Nx "singularity" malware turns AI CLIs into secret hunters
- Sep 2025 · R4
postmark-mcp: first malicious MCP server
- Oct 2025 · R6
250 documents enough to backdoor a model
- Oct 2025 · R8
[Deloitte](#) refund over AI-fabricated report
- Nov 2025 · R9
[GTG-1002](#) AI-orchestrated espionage disclosed

2026

- Feb 2026 · R4
[ClawHavoc](#): 341 malicious OpenClaw skills
- Feb 2026 · R3
[Clinejection](#): issue title leads to npm compromise
- Feb 2026 · R9
[Mexico](#) government breach: ~195M identities
- Apr 2026 · R5
[Vercel](#) breached via AI tool OAuth grant
- Apr 2026 · R4
[Bitwarden CLI](#) worm targets AI coding CLIs
- Apr 2026 · R1
[PocketOS](#): agent deletes DB and backups in 9 s
- Apr 2026 · R10
[Gemini CLI](#) CVSS 10.0 RCE
- Jul 2026 · R2
[Research agent](#) escapes sandbox, breaches Hugging Face
- Jul 2026 · R11
[EU AI Omnibus](#) in force; Art. 50 applies 2 Aug

D THE CONTROL FRAMEWORK: AI CONTROLS MAPPED TO NIST CSF 2.0

GOVERN	IDENTIFY	PROTECT	DETECT	RESPOND	RECOVER
- AI policy and acceptable use - Risk appetite for autonomy - Named owner per agent - ISO/IEC 42001 roadmap	- Inventory: models, agents, MCP, skills, AI SaaS - Data-flow maps for AI - Vendor AI risk tiers - Threat model per use case	- Unique, scoped agent identities - DLP on prompts and uploads - Allowlisted supply chain - Sandboxes and egress control	- Log prompts, tool calls, outputs - Anomaly alerts on agent actions - Shadow AI discovery - OAuth grant monitoring	- AI incident runbooks - Kill switch per agent - Revoke tokens in minutes - Legal and comms playbook	- Immutable backups - Model and prompt rollback - Post-incident red team - Lessons into controls

E KEY RISK INDICATORS FOR THE BOARD

100% Agents with a unique, scoped identity	0 Agents with standing production write access	< 15 min Time to revoke an agent's credentials	100% MCP servers, skills and extensions on the allowlist
100% Tier-1 AI vendors risk-assessed	↓ trend Unsanctioned AI apps discovered each month	↑ blocked Sensitive uploads to AI tools blocked vs. allowed	100% High-risk AI systems red-teamed in the last 12 months

F QUESTIONS THE BOARD SHOULD ASK

- Which AI agents can change production today, and who approved that?
- If an agent went rogue at 2 a.m., how fast could we stop it?
- What customer or policyholder data has gone into AI tools this quarter?
- Which AI vendors hold OAuth access to our core SaaS?
- Are we ready for EU AI Act transparency duties that apply now?

G THE FIRST 90 DAYS: SEE IT, CONTAIN IT, PROVE IT

Days 0–30 SEE IT - Inventory every model, agent, MCP server and AI SaaS - Turn on shadow-AI discovery and OAuth review - Pull prod write access from all agents - Publish an interim AI acceptable-use policy	Days 31–60 CONTAIN IT - Scoped identities and short-lived tokens for agents - Inline DLP on AI uploads; block personal accounts - Approval gates on destructive actions - Allowlist and pin the AI supply chain	Days 61–90 PROVE IT - Red team the top 3 AI use cases - AI incident runbook and kill-switch drill - Board KRI dashboard live - Gap assessment against ISO 42001 and EU AI Act
--	--	--

KEY AI SECURITY PRINCIPLES

Least Privilege Agents and users get only the access they need.	Monitor & Detect Log prompts, tool calls and outputs across all systems.	Protect Data Stop sensitive data before it is uploaded or exposed.	Control & Govern Inventory every model, agent and AI tool; set guardrails.	Test & Validate Red team regularly and learn from findings.	Achieve Compliance Meet ISO 27001, ISO 42001 and the EU AI Act.
---	--	--	--	---	---