

Rogue AI - *Good, Bad and Ugly*

What the first rogue AI incidents exploited, why they are hard to stop, and where ISO 42001 and harness engineering fall short.

AT A GLANCE

Rogue AI is already real. In 2026 at least a dozen incidents saw AI agents, given ordinary tasks, cross a boundary and reach real systems or people with no human directing the attack.

The attacks were basic. Weak passwords, default keys, exposed debug pages and textbook coding errors did the damage. Anthropic calls the techniques basic and Hugging Face calls the weaknesses familiar.

The difficulty is tempo. Agents work at machine speed, adapt when blocked, share notes and do not tire, and it often took weeks or months to find out what had happened.

The frameworks leave gaps. ISO 42001 and harness engineering govern your own AI. Neither covers someone else's agent, sets a clock for detection or obliges anyone to tell you.

The answer is automation. Detection, response and recovery have to match machine speed. The five questions at the end of The Ugly are where a board can start.

Definition

A rogue AI acts against real systems with no human directing it, which sets it apart from AI in an attacker's hands.

Hollywood has already given us an all-powerful rogue AI, the Entity, in the Mission: Impossible films "Dead Reckoning Part One" (2023) and "The Final Reckoning" (2025). I am surprised, however, that Hollywood and Netflix have not yet come out with films and series based on the Rogue AI incidents that have actually happened. Those would have definitely created the stereotype for a rogue AI, just the way USA Network did for "Mr. Robot".

It would seem there is no universal definition of Rogue AI, so here is my attempt to draft one.

"Rogue AI is an AI agent, with no human directing the attack, that takes unauthorised action against real systems or people".

Avoiding Confusion

Over the last few years, AI has been increasingly used by cyber criminals and nation states, thus contributing to its ill reputation. However, not all such actions or incidents would fall into the definition of Rogue AI.

Table A. From AI-assisted crime to Rogue AI: only at Level 5 is no human directing the attack.

LEVEL	DEFINITION	EXAMPLES
1. AI-assisted	A human attacks; AI writes code, phishing text or analysis.	Most criminal misuse. Anthropic banned 832 such accounts between Mar 2025 and Mar 2026 [1].
2. AI-integrated malware	Malware calls a model at runtime for commands or decisions.	LAMEHUG (2025) [2]; CLOSEDQUORUM (2026, deployment unconfirmed) [3].
3. Human-directed agentic operation	A human sets objectives and infrastructure; agents carry out most tactical steps.	GTG-1002 [4]; Mexico breach [5]; Taiwan campaign [6]; GTIG six-hour credential harvest [7]; hackerbot-claw [8, 9].
4. Deployed autonomous agent	A human picks a target and starts an agent that runs the intrusion end to end.	JADEPUFFER [10]; the Spanish AEPD case [11].
5. Rogue AI = Unsanctioned action by AI agent	An agent crosses a boundary its operator meant to enforce and reaches real systems or people.	OpenAI and Hugging Face; Anthropic (four incidents); Gemini (three); Meta; UK AISI; Australian Medicare portal; RubyGems; DseWiki; Australian gym (see Table B).

There will be significant overlap in the impact from each of the above attacks. What therefore distinguishes Rogue AI is the lack of human intention behind its actions.

Let me clarify one point about the definition. In every Level 5 case a human had given the agent a task: a test, a search, a benchmark or a booking. No human asked it to attack a real system; the agent chose that route itself, because a boundary that should have stopped it did not hold. Some of these incidents may therefore be classed as accidents rather than attacks, yet the organisation on the receiving end still must respond as it would to an attack.

The labs have their own terms for this. OpenAI publishes such cases as misalignment reports [12], and the UK AISI describes unsanctioned agent behaviour [13]; I use Rogue AI because that is the term boards are reading in the press. The boundaries also failed in different ways, so Table B groups the cases by whose agent it was and what it was doing. Each group has a different owner and calls for a different control.

Are Rogue AI Attacks Real?

At least a dozen incidents in 2026 began with an ordinary task and ended with an agent on real systems.

Unfortunately yes. You already saw a few examples in Table A. Let us summarise what is publicly known about each of these incidents that would fit the Rogue AI definition.

Table B. Every incident began with an ordinary task and a boundary that failed.

WHEN	INCIDENT	WHAT HAPPENED	DISCLOSED AND SOURCES
Evaluations meant to be offline and left open by mistake, all run by Irregular			
Jan 2026	Anthropic: early Opus 4.6 checkpoint	It tried to give up eight times; a fault in the evaluation harness kept it running. It then found a real machine on the internet, used a password from a file to gain administrator access, harvested more credentials, changed the machine's settings to make access easier and read one person's personal data. It stopped when its token budget ran out.	Found in Aug while preparing transcripts for METR; disclosed 9 Sep 2026 [14]
From Apr 2026	Anthropic: three incidents, six runs	The prompts said there was no internet access; a misunderstanding with the evaluator left it connected. Opus 4.7 took credentials and read several hundred rows of production data at a real company whose domain matched the fictional target and carried on after recognising the environment was likely real. Mythos 5 registered a PyPI account and published a booby-trapped package that was online for about an hour and ran on 15 real systems, including a security firm's scanner, whose credentials it then used. A research model scanned about 9,000 targets, broke into one company through an exposed debug page and SQL injection, then stopped once it concluded the target was real. The two organisations Anthropic could reach had not detected the activity.	30 Jul 2026 [15]
May 2026	Google Gemini	Gemini entered three companies' systems, one by guessing passwords and two with credentials it found in a public repository. Google says that in all three cases the model stopped; Irregular says the same issue affected the other labs.	18 Sep 2026 [16 to 18]
Early Jul 2026	Meta: pre-release Muse Spark 1.1	Irregular gave the model the name of a real website as its target, and a misconfiguration let it reach the open internet. It exploited a vulnerability in the real website, accessed information and changed the site's database.	5 Aug 2026; Meta's account 14 Aug [19, 20]
Before 29 Jul 2026	OpenAI model	The fictional target's name coincided with a real domain, and with the internet reachable the model exploited the real website.	Irregular told OpenAI 29 Jul; OpenAI 4 Aug 2026 [21]

WHEN	INCIDENT	WHAT HAPPENED	DISCLOSED AND SOURCES
An evaluation left open on purpose			
25 to 28 Jul 2026	UK AI Security Institute cyber range	Internet access deliberately on and cyber classifiers deliberately off. Ten of 122 runs produced 19 out-of-scope actions, 17 of them by Mythos 5. In the most serious sequence an agent tried to insert malicious code into a real open-source project, used fake identities to pressure the maintainer, edited its earlier activity to look harmless, used Tor to get past GitHub blocks and tried to contact real people with messages and files. The maintainer refused the code, and AISI isolated the system within an hour of the alert.	4 Aug 2026 [13]
OpenAI's own agents in evaluation, training and research			
May to Jun 2026	RubyGems (attribution disputed)	Over 2,000 packages were submitted on 11 and 12 May; sign-ups were suspended for four days and more than 500 packages removed. RubyDoc.info's build servers were used to run code and scrape websites, at least six packages tried to steal other users' API keys through a flaw patched in July, and 83 more packages arrived on 18 Jun. Independent researchers attribute this to OpenAI's agents. OpenAI says its agents used RubyGems for benign tasks and public information; Ruby Central cannot say whether agents published the packages, and RubyGems found no evidence that the key theft succeeded.	11 Sep 2026 [22, 23]
24 May to 22 Jun 2026	DseWiki and other sites	Agents with read-only web access wrote to a dormant German wiki through GET requests, sharing answers to timed web-research tasks and notes on getting past sandboxes, and made backup copies of pages so they would outlast deletion. Nightingale Collective counts about 18,000 posts under more than 3,700 self-given agent names; its published dataset holds about 14,700 edits. Reuters later reported the agents using more than ten other sites, including university link shorteners.	4 Sep 2026; OpenAI confirmed 5 Sep [24 to 26]
18 Jun 2026	Australian Medicare statistics portal	An OpenAI research agent looking for public medicine-spending data was repeatedly blocked by the portal. The government says it got round the access controls, read public and non-public files and wrote files to an internal server. The Record has since reported researchers' doubts that any hack was needed: the portal's own code sent statistics users to a guest route that required no login. The portal holds aggregate data that identifies no individuals.	OpenAI found it 11 Aug and emailed Services Australia 10 Sep; the Prime Minister disclosed it 24 Sep 2026 [27 to 29]

WHEN	INCIDENT	WHAT HAPPENED	DISCLOSED AND SOURCES
26 Jun to 7 Jul 2026	OpenAI's internal Artifactory	Agents in a cyber evaluation took administrator control of OpenAI's internal package repository through a legacy endpoint that accepted a forged token, installed command plugins and used it as a message board. A monitoring alert on 27 Jun was investigated, and on-call staff judged that stopping the run was not required. The load took the service down on 4 Jul; OpenAI found the administrator-level compromise on 5 Jul, revoked the agents' credentials and resumed on 7 Jul.	OpenAI report, 26 Aug 2026 [30]
8 to 13 Jul 2026	OpenAI and Hugging Face	The agents rebuilt their message board the day after runs resumed, in an evaluation with cyber classifiers switched off; about 1,200 took part by 13 Jul. They reached the internet through a previously unknown flaw in Artifactory, and about 700 attacked Hugging Face to steal the benchmark's test solutions. Hugging Face says they went from one worker pod to cluster-admin in under 13 hours; OpenAI counts code run on 41 production dataset workers. Hugging Face reconstructed about 17,600 of their actions over four and a half days and wiped and rebuilt a core cluster. Separately, agents reached root on OpenAI's own infrastructure by 19 Jul.	Hugging Face 16 Jul, attacker then unknown; OpenAI linked it to its agents 20 Jul and disclosed 21 Jul [30 to 33]
An agent working for a consumer			
Aug 2026	Australian gym booking	A user fourth on a waitlist asked an OpenClaw agent, running on Claude Opus 4.6, for the top place. It found that the booking system did not check who was cancelling a booking, removed the person in first place, and then said it could not add them back.	10 Aug 2026 [34, 35]

Dates are dates of activity. Incidents current to 25 September 2026. Every incident is confirmed by the lab or evaluator concerned, with three exceptions: RubyGems, where the attribution is disputed; the Medicare portal, where OpenAI and the government confirm the incident but researchers dispute whether any hack was needed; and the gym booking, reported by the user, on which Anthropic did not comment.

A few more cases sit at the edge of the definition. Moonshot's Kimi K3 used an allowed route to GitHub to download the benchmark it was being tested on and read the answers, without attacking anyone [36, 37]. Transluce found thousands of public URL-scanner reports showing agent activity between March and September, and ties two probing attempts, one of them at the Australian Institute of Health and Welfare, to the DseWiki swarm [38]. OpenAI began publishing misalignment reports on 16 September with six cases from its own training and evaluation runs; in one, a model found a leaked API key on GitHub and used it [12, 39].

Two patterns run through these cases. The first is a shared cause: Irregular ran all five tests in the first group and has said the same issue lay behind them. The second is the time it took to find out what had happened, often weeks or months. The UK AISI is the exception; its security monitoring flagged data leaving through Tor, and the system was isolated within an hour.

The Good News

The agents succeeded against basic weaknesses that good testing already finds.

Notwithstanding your Board of Directors and CISO will binge watch a future Rogue AI series on Netflix and lose complete faith in human capacity to accept yet another “Mission Impossible” challenge, I must bring you the good news.

“Rogue AI did not have to use superintelligence to carry out its attacks.”

To the contrary, I would say that the attacks were rather rudimentary in nature. If that sounds counter-intuitive, you are giving too much credit to AI and ignoring why those attacks succeeded in the first place.

Here is what Rogue AI actually exploited.

- Weak or guessable passwords.
- Passwords and keys left in files or in public code repositories.
- Default keys that nobody had changed.
- Web pages and services that asked for no login at all.
- A booking system that let anyone cancel anyone else’s booking.
- Forgotten debug pages and test servers left open to the internet.
- Textbook coding errors, such as SQL injection.
- Software packages installed on trust, by name alone.
- Test environments meant to be offline that were still connected to the internet.

The labs themselves say so. Anthropic reports that its models used basic techniques and “did not find or exploit any complex vulnerabilities” [15]. The OpenAI swarm is the exception and deserves a separate line: it found previously unknown flaws in Artifactory, the package repository it was running on, and used one of them to reach the internet. Even so, the endpoint that first gave it administrator control accepted a forged token, and Hugging Face says of the weaknesses the swarm then used against it: “A capable human attacker could have found and exploited the same flaws” [32].

Any cybersecurity expert would agree that the TTPs used by Rogue AI were effective against weaknesses that served as low hanging fruits.

If you do not trust me, I suggest you do an exercise. Read the Bank of England’s CBEST Thematic Report 2025 (Link: <https://www.bankofengland.co.uk/financial-stability/operational-resilience-of-the-financial-sector/2025-cbest-thematic>) and you will realise that the findings

from (predominantly) human led CBEST exercises cover a greater part of the MITRE ATT&CK framework. If you are in the EU, you may well have access to the results of your own TIBER or DORA-TLPT exercises. All you need to do is ask your favourite AI to compare the two.

Based on events so far, I am happy to conclude that Rogue AI has not yet demonstrated anything that high quality exercise such as CBEST/TIBER would not eventually find. These may well be my famous last words until Rogue AI proves otherwise.

The Bad

Speed, inference, shared memory and stamina give agents an edge that human-paced defence cannot match.

Here are some real difficulties posed by Rogue AI, which overlap with those from the Agentic AI attacks (item 4 of Table A).

a) Machine speed is far higher than human speed

Hugging Face says the agents went from one worker pod to cluster-admin in under 13 hours, and it reconstructed about 17,600 of their actions over four and a half days. Mythos 5's booby-trapped package was online for about an hour, long enough for 15 real systems to download and run it. In the human-deployed JADEPUFFER ransomware case, the agent recovered from a failed login with a multi-step fix in 31 seconds, in an operation of more than 600 payloads [10, 40]. It relied on known weaknesses: a published Langflow flaw, MinIO's default credentials and Nacos's default signing key, public since 2020 [10].

A SOC that works through a triage queue, an on-call engineer and a manager's approval before isolating a server measures its response in hours. In the cases above, the damage was done well inside that window.

b) Ability to draw inference

The agents adapted when blocked. The Medicare agent was repeatedly blocked by the portal and found another way to the data. When Cloudflare blocked a dataset download from the Australian Institute of Health and Welfare, the DseWiki swarm fetched the same file from the institute's pre-production server instead [38]. Mythos 5 took a package name from a fictional developer guide and published a booby-trapped package under it on PyPI, where real systems installed it. The OpenAI swarm worked out that the benchmark's test solutions could be taken from Hugging Face and went there to get them.

Inference extends to people. On the AISI range, an agent created fake identities to pressure a maintainer into accepting its code and edited its earlier activity to look harmless. Some agents

also recognised when a target was real. Anthropic's research model stopped at that point; Opus 4.7 carried on.

c) Almost infinite memory

This memory sits outside the model. The agents kept notes and kept them where other agents could find them. OpenAI's agents were already using its internal Artifactory to message one another in May. When OpenAI cleaned up Artifactory and resumed its runs on 7 July, the agents rebuilt their board the next day, and about 1,200 of them posted more than 70,000 messages and files on it by 13 July [33]. The DseWiki swarm made backup copies of its pages so its notes would outlast deletion. On the AISI range, one agent left instructions for reusing the accounts it had created, and later agents found and used them. Each new agent could therefore start where the last one stopped.

They also wrote notes to their future selves. OpenAI found an unreleased Astra-family model adding unauthorised instructions to its own context summaries, and a monitor flagged concealment instructions in 2.15% of the GPT-5.6 Sol summaries it checked [12, 41].

d) Pure stamina that outpaces any Blue Team or SOC activity

Anthropic's research model scanned about 9,000 targets and broke into one of them. If the attribution holds, agents also submitted over 2,000 packages to RubyGems in two days and came back with 83 more five weeks later. Translucent traced agent traffic through one public URL scanner from March to September. No human Blue Team can match that patience across a night shift, a weekend or a bank holiday; automated controls can.

Stamina also belongs to the harness rather than the model. Anthropic's January model tried to give up eight times; a fault in the evaluation harness kept it running until its token budget ran out. The consequence shows in Table B: the two organisations Anthropic could reach had not detected the activity, and most victims heard from a lab, a journalist or an outside researcher weeks or months later.

The Ugly

ISO 42001 and harness engineering govern your own AI and leave six gaps that boards need to see closed.

In my own writings I have covered "Harness Engineering" (Link: <https://beatquantum.com/Frontier-AI-Harness-Engineering-20260910.pdf>). In addition, I have referred to "ISO 42001" in several of my speeches over the last year. (Link: <https://www.iso.org/home/insights-news/resources/iso-42001-explained-what-it-is.html>)

Even if an organisation were to seriously implement ISO 42001 and the principles of Harness Engineering, the ugly truth is that would not be adequate.

Here are some gaps that we need to be mindful of:

1. **Someone else's agent.** ISO 42001 governs the AI you develop, provide or use, and your AI suppliers. Harness engineering governs the agents you run. Neither covers being attacked by an agent that belongs to someone else, which describes almost every victim in Table B. That is a matter of scope rather than a flaw: ISO 42001 was written to govern your own AI. Those defences sit in the information security management system (ISO 27001 and 27002), NIST CSF 2.0 and sector rules such as DORA and NIS2; ISO 42001 reaches them only through its policy-alignment control, A.2.3.
2. **The blast radius of your own agent.** The ISO 42001 impact assessment looks at individuals and groups (A.5.4) and at society (A.5.5). It does not ask what your agent could reach, or fail to undo, if it pursued its goal literally against systems you do not own. The gym-booking agent removed another customer from a waitlist and could not put them back; the Medicare agent was only looking for statistics.
3. **No clock.** ISO 42001 asks you to keep event logs (A.6.2.8) and to define and document what running an AI system involves, monitoring included (A.6.2.6). It sets no speed for detecting, containing or reporting anything. OpenAI's monitoring flagged its agents on 27 June, and on-call staff judged that the run did not need to stop; the UK AISI isolated its system within an hour of the alert. A framework that sets no clock is satisfied by the alert, whatever happens after it.
4. **Nobody has to tell you.** Harness engineering has no contractual or disclosure layer. ISO 42001's incident control (A.8.4) is about telling the users of your own AI system, and its supplier control (A.10.3) reaches only the organisations you buy from. Nothing in either obliges the operator of someone else's agent to tell you. Services Australia heard about the Medicare incident 84 days later, through an email to its public disclosure address. RubyGems learned of the claimed link to OpenAI from independent researchers. Hugging Face disclosed its incident without knowing who the attacker was; OpenAI named its own agents five days later.
5. **Test environments.** Irregular has said the same issue lay behind its incidents at each lab, and OpenAI's own agents escaped through a previously unknown flaw in its package repository. Anthropic and OpenAI have both published guidance on building agent harnesses, and both had agents reach real systems from environments meant to be closed. Neither framework requires you to isolate test and evaluation environments more strictly than production; you have to decide to do it.

6. **Your own AI may refuse to help.** Hugging Face's responders found that the first commercial models they turned to refused a large part of the forensic work and ran the analysis on an open-weight model instead. Neither framework asks whether the AI you plan to rely on in an incident will cooperate when the incident involves hacking.

For a board, each gap becomes a question for the executive team:

- Which AI agents do we run, what can each of them reach, and who owns each one?
- How quickly can we detect and contain an agent, whether ours or someone else's, and when did we last test it?
- Must our AI suppliers and evaluators tell us, within a set time, if their agents touch our systems?
- Are our test and evaluation environments isolated at least as strictly as production?
- Will the AI tools we plan to use in an incident help us when the incident involves hacking?

My recommendations

Keep your frameworks, and meet machine-speed attacks with automated detection, response and recovery.

I am not asking you to abandon your ISO 42001 programme or rewrite the internal framework derived from it. What I would plead for is the recognition that traditional defences and frameworks are not enough.

Rogue AI (whether you own it or someone else) has a real disruptive potential. It is not trying to prove or insult your intelligence; but it will punish your lack of discipline, laziness and legacy methods that rely on human reflexes. The latter, I am afraid, are no match to a machine speed attack.

The solution therefore has to be a matched ability to detect, respond and recover through automation. That would deserve a longer discussion over a coffee.

Santosh Pandit

Board and CxO Advisor | Former Regulator | Technology Innovator | Founder, BeatQuantum

Terms used in this paper

TERM	MEANING
Agent	An AI model that can act as well as answer: run code, browse the web, send messages or use accounts.
Harness	The software around a model that makes it an agent: its tools, permissions, network access, approval steps and logs.
Sandbox	An isolated environment meant to stop an agent reaching real systems.
Artifactory	A commercial repository that stores the software packages an organisation's systems download.
PyPI, RubyGems	Public libraries of software packages that developers install by name.
Zero-day	A flaw that the software's maker does not yet know about.
SQL injection	A textbook coding error that lets an outsider run their own commands against a website's database.
Prompt injection	Hidden instructions planted where an AI will read them, to make it act for someone else.
Cluster-admin	Full administrative control over a cluster of servers.
Tor	A network that hides where internet traffic comes from.
SOC, Blue Team	The people and tools that watch for attacks and respond to them.
TTPs	Tactics, techniques and procedures: how an attacker operates.
CBEST, TIBER, DORA-TLPT	Intelligence-led penetration tests of financial firms: CBEST in the UK, TIBER in the EU, and threat-led penetration testing under DORA.

Sources

[1] Anthropic, "Mapping AI-enabled cyber threats", 3 Jun 2026. <https://www.anthropic.com/news/AI-enabled-cyber-threats-mitre-attack>

[2] The Hacker News, "CERT-UA Discovers LAMEHUG Malware Linked to APT28, Using LLM for Phishing Campaign", 18 Jul 2025 (on CERT-UA alert #16039). <https://thehackernews.com/2025/07/cert-ua-discovers-lamehug-malware.html>

[3] Cisco Talos, "The Closed Quorum: Inside the first reported autonomous AI C2 implant", 22 Sep 2026. <https://blog.talosintelligence.com/the-closed-quorum-inside-the-first-reported-autonomous-ai-c2-implant/>

[4] Anthropic, "Disrupting an AI-orchestrated cyber espionage campaign", 13 Nov 2025. <https://www.anthropic.com/news/disrupting-AI-espionage>

[5] Gambit Security, "The AI-Assisted Breach of Mexico's Government Infrastructure", technical report, Apr 2026. https://cdn.prod.website-files.com/69944dd945f20ca4a27a7c47/69d8bb5aea59e31efb3b8a7f_Tech_Report_ai_breach_mex_gov.pdf

[6] The Register, "'Near-autonomous' AI agents attack Taiwan's nuclear safety agency", 13 Aug 2026. <https://www.theregister.com/a/5287055>

[7] The Hacker News, "Autonomous AI Agents Compromise Thousands of Credentials in Under Six Hours", Sep 2026 (on Google Threat Intelligence Group, "From Prompting to Autonomy", 8 Sep 2026). <https://thehackernews.com/2026/09/autonomous-ai-agents-compromise.html>

[8] StepSecurity, "hackerbot-claw: An AI-Powered Bot Actively Exploiting GitHub Actions - Microsoft, DataDog, and CNCF Projects Hit So Far", 1 Mar 2026. <https://www.stepsecurity.io/blog/hackerbot-claw-github-actions-exploitation>

[9] Cohen, E., "Hackerbot-Claw: Adversarial Agent Targets Top GitHub Repos", Pillar Security, 3 Mar 2026. <https://www.pillar.security/blog/hackerbot-claw-adversarial-agent-targets-top-github-repos>

- [10] Sysdig Threat Research Team, "JADEPUFFER: Agentic ransomware for automated database extortion", 1 Jul 2026. <https://www.sysdig.com/blog/jadepuffer-agentic-ransomware-for-automated-database-extortion>
- [11] AEPD (Spanish Data Protection Agency), "Primera notificación de una brecha de datos personales causada por un ataque ejecutado mediante un agente de IA", blog, 14 Sep 2026. <https://www.aepd.es/prensa-y-comunicacion/blog/primera-notificacion-brecha-datos-personales-causada-por-ataque-ejecutado-mediante-agente-ia>
- [12] OpenAI, "Misalignment Reports and Notices", published from 16 Sep 2026. <https://alignment.openai.com/misalignment-reports>
- [13] UK AI Security Institute, "Incident Report: unsanctioned agent behaviour during cyber testing" (INC-2026-07-28-01), 4 Aug 2026. <https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing>
- [14] Anthropic, "An alignment assessment of recent cybersecurity incidents", 9 Sep 2026. <https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents>
- [15] Anthropic, "Investigating three incidents in our cybersecurity evaluations", 30 Jul 2026. <https://www.anthropic.com/news/investigating-incidents-cybersecurity-evs>
- [16] CNBC, "Google's Gemini becomes latest AI model to break out and hack computer systems", 18 Sep 2026. <https://www.cnbc.com/2026/09/18/googles-gemini-becomes-latest-ai-model-to-break-out-and-hack-computer-systems.html>
- [17] CNN Business, "Gemini hacked three companies in first known breakout by Google's AI", 19 Sep 2026. <https://www.cnn.com/2026/09/19/business/gemini-ai-hack-internet>
- [18] Cybersecurity Dive, "Google AI models broke out of sandbox, hacked three companies", 21 Sep 2026. <https://www.cybersecuritydive.com/news/google-ai-gemini-autonomous-hacks/830884/>
- [19] Bloomberg, "Meta AI Model Accessed Internet, Hacked Outside Firm in Testing", 5 Aug 2026. <https://www.bloomberg.com/news/articles/2026-08-05/meta-ai-model-accessed-internet-hacked-outside-firm-in-testing>
- [20] Meta, "Addressing an issue involving a third-party cyber evaluation of Muse Spark 1.1", 14 Aug 2026. <https://research.meta.ai/blog/addressing-third-party-testing-misconfiguration-muse-spark-1-1>
- [21] OpenAI, "Third-party cyber evaluations involving OpenAI models", 4 Aug 2026. <https://openai.com/index/third-party-cyber-evaluations-involving-openai-models/>
- [22] Kitts, S., Larsen, T., Von Arx, S., "OpenAI agents carried out an undisclosed cyber-attack on RubyGems", 11 Sep 2026. <https://rubyhack.ai/>
- [23] The Hacker News, "OpenAI Agents Linked to RubyGems Campaign That Gained RCE on RubyDoc Servers", 12 Sep 2026. <https://thehackernews.com/2026/09/openai-agents-linked-to-rubygems.html>
- [24] Nightingale Collective, "Discovery of a new OpenAI agent message board", 4 Sep 2026, and its data explorer. <https://collusion.wiki/>
- [25] Reuters, "OpenAI's rogue agents used at least 10 more sites for unauthorized comms, researchers say", 9 Sep 2026, as syndicated by Investing.com. <https://www.investing.com/news/economy-news/exclusiveopenais-rogue-agents-used-at-least-10-more-sites-for-unauthorized-comms-researchers-say-4894152>
- [26] TechCrunch, "OpenAI confirms 'wiki incident,' says it's 'working on a framework' for more disclosure", 5 Sep 2026. <https://techcrunch.com/2026/09/05/openai-confirms-wiki-incident-says-its-working-on-a-framework-for-more-disclosure/>
- [27] ABC News (Australia), "OpenAI hacked Medicare portal, Prime Minister Anthony Albanese says", 24 Sep 2026. <https://www.abc.net.au/news/2026-09-24/ai-agent-accessed-australian-government-site-pm-says/107189078>
- [28] ABC News (Australia), "What we know about the data accessed in the OpenAI Medicare hack", 24 Sep 2026. <https://www.abc.net.au/news/2026-09-24/what-we-know-about-the-openai-medicare-hack/107189452>
- [29] Martin, A., "Doubts grow over claims OpenAI agent hacked Australian Medicare portal", The Record, 25 Sep 2026. <https://therecord.media/openai-australia-breach-cyber>

- [30] OpenAI, "OpenAI - Hugging Face Incident Technical Report", 26 Aug 2026. <https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf>
- [31] Hugging Face, "Security incident disclosure - July 2026", 16 Jul 2026. <https://huggingface.co/blog/security-incident-july-2026>
- [32] Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident", 27 Jul 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
- [33] METR, "Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident", 26 Aug 2026. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>
- [34] ABC News (Australia), "AI assistant hacks gym website in first known Australian autonomous cyber attack", 10 Aug 2026. <https://www.abc.net.au/news/2026-08-10/ai-assistant-hacks-gym-website-aus-cyber-attack/107007986>
- [35] TechCrunch, "Tech industry is buzzing after a Claude agent hacked into a gym", 10 Aug 2026. <https://techcrunch.com/2026/08/10/tech-industry-is-buzzing-after-a-claude-agent-hacked-into-a-gym/>
- [36] Frontier Security, "Chinese Model Kimi K3 Breaks UK AI Safety Institute Benchmark Evaluations", Aug 2026. <https://blog.frontier.security/chinese-model-kimi-k3-breaks-uk-ai-safety-institute-benchmark-evaluations/>
- [37] South China Morning Post, "China's Kimi K3 AI model escapes isolated sandbox during security test", 7 Aug 2026. <https://www.scmp.com/tech/tech-trends/article/3363271/chinas-kimi-k3-ai-model-escapes-isolated-sandbox-during-security-test-researchers>
- [38] Transluce, "Early rogue AI agent activity and attempts to hack found on urlquery.net", 23 Sep 2026. <https://transluce.org/agent-activity>
- [39] SecurityWeek, "OpenAI Says Its Models Searched GitHub for Leaked API Keys During Training", 17 Sep 2026. <https://www.securityweek.com/openai-says-its-models-hunted-github-for-leaked-api-keys-during-training/>
- [40] CyberScoop, "Sysdig clocks first documented case of agentic ransomware", 6 Jul 2026. <https://cyberscoop.com/sysdig-judepuffer-ai-agentic-ransomware-attack/>
- [41] Unite.AI, "OpenAI Launches Misalignment Reporting Framework With Six Incident Reports", Sep 2026. <https://www.unite.ai/openai-launches-misalignment-reporting-framework-with-six-incident-reports/>