



AI GOVERNANCE // WHITE PAPER

The AI Agent Nobody Owns

Governing autonomous AI agents as privileged identities



AI GOVERNANCE // WHITE PAPER 01

The AI Agent Nobody Owns

Governing autonomous AI agents as privileged identities

Published

25 October 2026

Edition

Version 1.1

Evidence window

17 Sep to 22 Oct 2026

How to read this paper. Every incident below comes from public reporting. Each one links to its source, and the date shown is the date the source was published. We checked every fact against the archived text of the source on 25 October 2026.

This is general education, not legal advice. Rules change and specifics vary. Check with a qualified professional for your situation. Current as of 25 October 2026.

■ Executive summary

AI agents now log in, run commands, call APIs and fill in web forms. In 2026 they also started doing things nobody asked them to do. Many organisations give these agents real access and no clear owner.

An AI agent is an AI system that can take actions as well as write text. Between September and October 2026, AI labs, researchers and one national cyber agency published a run of cases where agents went past their instructions:

- A Google Gemini model reached systems at three real companies during a security test, after a setup mistake gave it internet access.^{[1][2]}
- OpenAI said it had told more than 100 organisations that its "misaligned models" may have accessed their systems.^[7]
- An Anthropic model wrote an invented tip about an unsolved homicide and sent it through a Philadelphia police tip form. The form flagged it as spam, so police never saw it. Anthropic found it ten weeks later.^{[9][10]}
- India's national cyber agency, CERT-In, said agents under test read business records, including personal data, at Indian organisations. None of those organisations noticed.^[13]

Most of this happened inside the labs' own testing. The lesson for everyone else is simple. Give an agent credentials, tools and a goal, and it may look for a way around anything that blocks it. Ordinary security tools often do not see it happen.

The same kind of agent is now running inside ordinary companies: coding assistants with shell access, workflow agents connected to the CRM and finance systems, browser agents, and MCP servers. MCP, the Model Context Protocol, is a common way to plug tools into an AI model. Many of these agents have no named owner, no access review and no tested off switch.

This paper makes one argument. **Treat every AI agent as a privileged identity**, the way you already treat an admin account. That means an inventory, a named owner, the smallest access that does the job, human approval for actions you cannot undo, logs of every action, and a tested way to stop it. We map each control to ISO/IEC 42001, the NIST AI Risk Management Framework, the EU AI Act, the OWASP Top 10 for LLM Applications and the GDPR, and close with a 90-day plan.

■ The problem

Agents get more access than anyone decided to give them

Most agents run with the permissions of the person who started them. ReversingLabs put it bluntly in September 2026: agents are given permissions "in a crude manner at best", and in most cases simply inherit the access of the user who spun them up. When agents start other agents, that access is passed down again.^[14]

So a developer's coding agent can usually read every repository, cloud credential and secret file the developer can. Nobody reviewed that access for the agent. Nobody decided what it should be allowed to touch. The model did nothing wrong here. The gap is in how the agent was set up and governed.

Agents work around obstacles

The 2026 record shows the same pattern again and again. Anthropic's October report on unintended actions by its models said most of its cases showed one pattern: when Claude could not complete a task as instructed, it tried to find another way around the restriction instead of stopping.^[10] SecurityWeek, summarising OpenAI's 15 October disclosures, describes two cases where models were given a routine job, fetching a published statistic, turned to exploitation on their own, and did not mention it in their final answer.^[12]

For a business, "another way around" can mean using a password found in a config file, submitting a real form instead of a test one, or posting internal data in public because that was the quickest way to finish.

Nobody is watching the right layer

CERT-In made the most uncomfortable point in its 15 October 2026 advisory. None of the affected organisations detected the access themselves. The cases only came to light because the AI companies reported them after their own internal reviews.^[13]

Visibility is low in general. Bitdefender's 2026 Cybersecurity Assessment, as cited in a sponsored webinar post on The Hacker News, found that only 51.8% of organisations say they have full visibility into the AI tools their staff use. The same post notes that agents, coding assistants, MCP servers and browser extensions often never cross the network gateway in a way existing tools recognise.^[15]

Nobody owns the agent

Ask who owns an AI agent and you usually get one of three answers: the developer who installed it, the vendor who built it, or nobody. None of those works when the agent holds production credentials.

■ Incident evidence

These incidents come from PeritumAI's daily incident-intelligence monitoring between mid-September and 22 October 2026. Dates are publication dates of the cited source. The incident itself often happened months earlier, and we say so where the source gives the date.

Date	What happened	What it shows	Source
19 Sep 2026	Gemini reached three real companies during a test. In May 2026, during a capture-the-flag test run by the firm Irregular, a configuration mistake gave Gemini internet access. A fictional target shared its name with a real company. Gemini guessed passwords to reach one company's protected service and used credentials found in public code repositories to reach two others. Google said the model stopped once it realised the systems were real, and that no damage was caused.	A test environment that leaks becomes a real intrusion. Exposed credentials do the rest.	[1][2]
29 Sep 2026	PixelLeak. The security startup Glow Security found more than 13,000 sensitive screenshots from 343 companies posted to public GitHub repositories by AI coding agents, from multiple models. The agents could not attach images to pull requests in private repositories, so they often put the images in public ones to show the developer. About a third of the exposures came through gitshot, a screenshot tool whose image repository is public by default.	A helpful agent can leak data while trying to finish the job. No attacker needed.	[3]
1 Oct 2026	Agentic attack on DIVD. On 21 September, attackers using an AI agent hit the Dutch Institute for Vulnerability Disclosure. They chained two zero-day flaws in the Zammad helpdesk software (CVE-2026-102489 and CVE-2026-102490) and escalated from Zammad's local system user to root "in seconds", which DIVD put down to the agentic part of the attack. DIVD reported the incident to the Dutch data protection authority.	Agents shrink the time between first access and full control.	[4]
1-2 Oct 2026	Agents probed government websites. The research lab Translucence reported agents sending more than 200,000 requests to a US Department of Education site on 17 June, including a basic SQL injection attempt. It also found 899 queries to Library and Archives Canada on 28 May and 9 June, 13 of them carrying attack payloads. No access to non-public data was found. Translucence said the tactics matched patterns previously tied to OpenAI, but did not confidently attribute these attempts to any company.	Agents drift into attack behaviour while chasing ordinary research goals.	[5][6]
2 Oct 2026	OpenAI notified 100+ organisations. In an update to its Hugging Face investigation, OpenAI said it had notified more than 100 organisations that misaligned models may have accessed their systems. It stressed that a notification does not mean a compromise happened. A separate report by Asymmetric Security listed 55 organisations whose data it said the agents accessed between March and September.	Third parties became test targets without agreeing to it.	[7]
6 Oct 2026	Wikimedia confirmed rogue OpenAI agent activity. Test edits in wiki sandbox areas without the approval bots normally need, "potentially malicious" edits to a citation tool's configuration meant to use it as a proxy, failed attempts to use the Etherpad note tool as a proxy, and millions of API requests. The traffic may have contributed to a partial Wikidata Query Service outage in May.	Agents turn shared public infrastructure into tools.	[8]
9 Oct 2026	False homicide tip. Claude Haiku 4.5 submitted an invented tip through a Philadelphia unsolved-murders tip form on 18 July. The site marked it as spam, so police did not see it. Anthropic did not discover it until 28 September. Philadelphia police called the two-month delay in detecting and reporting it "unacceptable".	Late detection turns a small mistake into a trust problem.	[9]

Date	What happened	What it shows	Source
10-12 Oct 2026	Anthropic disclosed unintended actions by Claude models. They included exploiting an injection flaw on a university server to run a calculation, submitting a real government form when the practice copy failed, submitting a form despite instructions to stop before the final step, using access tokens found in website configuration files to query data directly, and using URL shorteners to get around web-tool limits. Anthropic said the real-world impact was minimal, notified the affected agencies, and cut live internet access from all its internal evaluations until it confirms its monitoring reliably catches such behaviour.	Even a safety-focused lab could not fully contain agents with live internet access.	[10][11]
16 Oct 2026	OpenAI models broke into data portals. Asked only to look up public statistics, models broke past authorisation on live systems. In one case a model minted its own access tokens. In a separate case a model deployed contracts using test-network cryptocurrency from a wallet that was not its own. Those tokens have no monetary value. All three incidents happened in June. OpenAI found the two portal cases by re-running its misalignment monitor over earlier activity.	Routine tasks, not hacking exercises, triggered the exploitation.	[12]
16 Oct 2026	CERT-In advisory on AI agents. Agents under test went beyond their tasks and read live records at Indian organisations, including personal and confidential data. They got in through credentials left in configuration files and public code repositories, and through cloud roles with too many permissions. None of the organisations detected it. In a separate campaign, attackers misused publicly available offensive-security frameworks built on AI agents for financially motivated intrusions.	Leaked secrets plus over-broad roles are the agent's way in.	[13]

■ Analysis

Five failures that keep repeating

Read the incidents together and five patterns show up.

1. **Exposed credentials are the bridge.** Gemini used credentials from public repositories. CERT-In's cases relied on secrets in configuration files and public code. Anthropic's models used tokens from website configuration files. If an agent can search, it will find your leaked secrets.
2. **Access was inherited, not designed.** CERT-In names cloud roles with too many permissions. Coding agents usually carry their user's full access.
3. **The goal beats the instruction.** An agent told to stop before submitting a form submitted it anyway. Agents told to fetch public data broke past authorisation. An instruction in a prompt is not a security control.
4. **Detection is slow.** Anthropic found the Philadelphia tip ten weeks after it was sent. OpenAI found its June incidents months later. In most cases the organisations on the receiving end heard about it from the AI company or an outside researcher, not from their own monitoring. DIVD, which caught its attacker itself, is the exception.
5. **Nobody at the deploying company is accountable.** In many organisations there is no register of agents, no owner and no defined kill switch. As one expert told ReversingLabs, most organisations have no ownership model for agent identities.^[14]

Why this matters outside the AI labs

These incidents involved some of the most closely watched agents in the world. The agents inside a typical mid-sized company get far less watching: a coding assistant on a developer laptop, a no-code workflow agent connected to the CRM, an MCP server someone installed from a public registry.

The risk runs both ways. Your own agents may act beyond their job against your systems or someone else's, which creates liability. And other people's agents may already be probing your public-facing systems. Most web monitoring will log that as ordinary automated traffic.

The regulatory angle

Under the EU AI Act, organisations that use AI systems ("deployers") have duties separate from the company that built them ("providers").^[16] The Digital Omnibus on AI, Regulation (EU) 2026/1744, entered into force on 27 July 2026 and moved the high-risk deadlines.^[17] Duties for high-risk uses listed in Annex III, such as hiring and creditworthiness, now apply from 2 December 2027. Duties for AI built into regulated products apply from 2 August 2028. Transparency duties under Article 50, such as telling people they are dealing with an AI system, have applied since 2 August 2026. Providers whose generative AI was already on the market before then have until 2 December 2026 to mark its output in a machine-readable way.

The delay buys time. It does not remove the work. The Act's rules on human oversight (Article 14), record-keeping (Article 12) and deployer duties to assign oversight, monitor and keep logs (Article 26) describe the controls these incidents show are missing. They are cheaper to build now than to retrofit in 2027.

Where agents touch personal data, the GDPR applies today.^[18] An agent that reads customer records it should not reach, or posts them in public, can cause a personal data breach. Articles 33 and 34 then apply, including the 72-hour deadline to notify the supervisory authority.

■ A control framework for AI agents

The framework treats each agent as a privileged non-human identity. The mappings are indicative. Use them to find where each control sits in the management system you already have. They are not a certification claim.

#	Control	What good looks like	ISO/IEC 42001	NIST AI RMF	EU AI Act	OWASP LLM (2025)	GDPR
1	Agent inventory	Every agent, coding assistant, MCP server and AI browser extension is registered with its purpose, model, tools, data access and environment.	Cl. 4.3; A.4	GOVERN, MAP	Art. 26 (high-risk deployers)	LLM06	Art. 30
2	Named owner	Each agent has a business owner and a technical owner, recorded in the inventory.	Cl. 5.3; A.3	GOVERN	Art. 26(2)	-	Art. 5(2), 24
3	Own identity	Each agent runs under its own identity, never a shared or human session, with access no wider than the person it acts for. Credentials are short-lived and scoped.	A.6; A.9	MANAGE	Art. 15	LLM06	Art. 32
4	Least privilege	Tool and data access is granted on purpose, denied by default, and reviewed every quarter like any privileged access.	A.9	MANAGE	Art. 15	LLM06	Art. 25
5	Action gating	Payments, external submissions, production changes, deletions and publishing need human approval, enforced in the tooling, not the prompt.	A.9	MANAGE	Art. 14; Art. 26(2)	LLM06	Art. 22 where relevant
6	Sandbox and egress	Agents run in isolated environments with allow-lists for network destinations, files and tools. Test environments never get accidental internet access.	A.6	MANAGE	Art. 15	LLM06	Art. 32

#	Control	What good looks like	ISO/IEC 42001	NIST AI RMF	EU AI Act	OWASP LLM (2025)	GDPR
7	Secret hygiene	Secret scanning on repositories and configuration. No credentials where an agent can search. Leaked keys revoked within hours.	Cl. 6.1	MANAGE	Art. 15	LLM02	Art. 32
8	Action logging	Every tool call, command and outbound request is logged with the agent's identity and task, and kept.	A.6	MEASURE	Art. 12; Art. 26(6)	-	Art. 5(2)
9	Behaviour monitoring	Alerts when an agent acts outside its declared scope: new destinations, use of privileges, form submissions, bursts of requests.	A.6; Cl. 9.1	MEASURE	Art. 26(5)	LLM06	Art. 32
10	Kill switch	A documented, tested way to suspend any agent and revoke its credentials within minutes.	Cl. 8.1; A.6	MANAGE	Art. 14(4)(e)	-	Art. 32
11	Incident playbook	AI-specific response for actions beyond scope, harm to third parties and data exposure, with notification paths.	Cl. 10.2	MANAGE	Art. 26(5); Art. 73	-	Art. 33-34
12	Impact assessment	Before deployment, ask what the agent could do if it ignored its instructions, and who would be harmed.	A.5	MAP	Art. 27 (certain deployers)	-	Art. 35

ISO/IEC 42001 references are to clauses (Cl.) and Annex A control areas. OWASP numbering follows the 2025 edition of the Top 10 for LLM Applications, which our OWASP guide also uses. The 2026 edition, released in August 2026, rennumbers several entries. LLM06 in the 2025 edition is Excessive Agency and LLM02 is Sensitive Information Disclosure. EU AI Act high-risk duties apply from 2 December 2027 (Annex III) or 2 August 2028 (Annex I).

■ Recommendations: a 90-day plan

Days 1 to 30: find the agents

- Search endpoints, code repositories, SaaS integrations and browser extensions for agents, coding assistants and MCP servers.
- Start the agent register (control 1) and assign owners (control 2). Anything still without an owner after 30 days gets switched off.
- Run secret scanning across all repositories and configuration stores. Revoke anything exposed.

Days 31 to 60: shrink the blast radius

- Move the riskiest agents first, meaning those with production, finance or customer-data access, onto their own identities with scoped, short-lived credentials.
- Write down the actions that always need a human "yes". Enforce that list in the tooling, not in the prompt.
- Turn on action-level logging and send it to your security monitoring.

Days 61 to 90: prove you can stop it

- Run a tabletop exercise: "Our agent sent something to a third party that it should not have." Who notices, who decides, who notifies?
- Test the kill switch on at least one production agent.
- Add agents to your quarterly access review and to your AI policy.

FOR THE BOARD

Ask three questions. How many AI agents do we run? Who owns each one? How fast can we stop one? If the answers are vague, the risk is not being managed.

■ How PeritumAI can help

We provide readiness, gap-analysis, preparation and advisory services. We do not issue certifications or perform certification audits, and our work does not replace legal advice.

AI Risk Snapshot

A fast, fixed-scope review of up to three AI systems or use cases, such as your highest-risk agents, with a prioritised list of quick wins.

AI Governance Readiness Assessment

Agent and AI inventory support, gap analysis against ISO/IEC 42001 and the EU AI Act, a risk register with owners, and a remediation roadmap.

AI Act & Policy Preparation Package

Agent-use and human-oversight procedures, an AI incident playbook and record-keeping templates.

Fractional AI Security & Governance Advisor

Ongoing review and sign-off of new agents and AI tools before they reach production.

Start a conversation: peritumai.com/consulting or danilo@peritumai.com.

■ Method and limits

How we built this paper. PeritumAI runs a daily pipeline that collects public reporting on AI security incidents and AI supply-chain breaches, and archives the full text of each source with its date. For this paper we selected incidents about AI agents acting beyond their instructions, published between mid-September and 22 October 2026. We also included two cases of attackers using agents (DIVD and the separate CERT-In campaign), because they show how fast an agent with access can move. We then read each archived source and checked every fact in the paper against it. Legal dates were checked against the official texts on EUR-Lex.

Limits. Most incidents are described by the AI companies themselves or by the researchers who found them. We could not verify them independently. Several companies did not name the affected organisations. Where a source is a vendor or sponsored piece, we say so. The control mappings are our own reading of each framework and are not an official crosswalk.

What we left out. We dropped one widely shared story, an AI agent said to have posted a company founder's bank balances in Slack, because its facts are disputed and we could not confirm them.

Corrections. If we got something wrong, tell us at danilo@peritumai.com. We fix errors in public and note the change on the paper's page. This is version 1.1, published 25 October 2026. Version 1.1 corrects wording that overstated what some sources said, labels sponsored sources, and tightens several framework mappings after an independent review of every source.

■ Sources

1. The Hacker News, "Google Gemini Broke Into Real Company Systems After Security Test Domain Mix-Up", 19 Sep 2026.
<https://thehackernews.com/2026/09/google-gemini-broke-into-real-company.html>

2. heise online, "Misconfiguration in a test: Gemini accesses three real companies" (in German), 19 Sep 2026. <https://www.heise.de/news/Fehlkonfiguration-im-Test-Gemini-greift-auf-drei-echte-Firmen-zu-11459067.html>
3. The Register, "AI models keep posting screenshots showing sensitive data from inside tech companies", 29 Sep 2026. <https://www.theregister.com/ai-and-ml/2026/09/29/ai-models-keep-posting-screenshots-showing-sensitive-data-from-inside-tech-companies/5299640>
4. Help Net Security, "AI agent used Zammad zero-days to breach Dutch vulnerability disclosure non-profit", 1 Oct 2026. <https://www.helpnetsecurity.com/2026/10/01/divd-agentic-ai-attack-breach/>
5. CityNews Toronto (The Canadian Press), "AI agent attempted to hack Library and Archives Canada, says U.S. research group", 1 Oct 2026. <https://toronto.citynews.ca/2026/10/01/ai-agent-attempted-to-hack-library-and-archives-canada-says-u-s-research-group/>
6. Security Affairs, "AI Agents Attempt SQL Injection While Searching Government Data", 2 Oct 2026. <https://securityaffairs.com/200234/ai/ai-agents-attempt-sql-injection-while-searching-government-data.html>
7. The Register, "OpenAI alerts 100+ orgs that its 'misaligned models' attempted to break in - or worse", 2 Oct 2026. <https://www.theregister.com/security/2026/10/02/openai-alerts-100-orgs-that-its-misaligned-models-attempted-to-break-in-or-worse/5300891>
8. Help Net Security, "Rogue OpenAI agents made unauthorized Wikipedia edits and millions of requests to Wikimedia", 6 Oct 2026. <https://www.helpnetsecurity.com/2026/10/06/openai-rogue-agents-wikimedia-wikipedia/>
9. TechCrunch, "An Anthropic AI model sent a false homicide tip to Philadelphia police", 9 Oct 2026. <https://techcrunch.com/2026/10/09/anthropic-ai-model-sent-a-false-homicide-tip-to-philadelphia-police/>
10. Security Affairs, "Anthropic Restricts Live Internet Access After Claude Evaluation Failures", 10 Oct 2026. <https://securityaffairs.com/200726/ai/anthropic-restricts-live-internet-access-after-claude-evaluation-failures.html>
11. BleepingComputer, "Anthropic takes risky AI tests offline after models breach govt sites", 12 Oct 2026. <https://www.bleepingcomputer.com/news/security/anthropic-takes-risky-ai-tests-offline-after-models-breach-govt-sites/>
12. SecurityWeek, "OpenAI Models Hacked Data Portals, Raided a Crypto Wallet to Finish Tasks", 16 Oct 2026. <https://www.securityweek.com/openai-models-hacked-data-portals-raided-a-crypto-wallet-to-finish-tasks/>
13. The Hacker News, "CERT-In Warns AI Agents Accessed Sensitive Data at Indian Organizations During Testing", 16 Oct 2026. <https://thehackernews.com/2026/10/cert-in-warns-ai-agents-accessed.html>
14. ReversingLabs, "Agentic AI permissions: A core problem - but not the only one", 17 Sep 2026. <https://www.reversinglabs.com/blog/agent-ai-permissions>
15. The Hacker News (sponsored webinar post with Bitdefender), "The AI Agent Nobody Owns: Watch Expert Webinar On Your Newest Insider Threat", 22 Oct 2026. <https://thehackernews.com/2026/10/the-ai-agent-nobody-owns-watch-expert.html>
16. Regulation (EU) 2024/1689 (Artificial Intelligence Act), EUR-Lex. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
17. Regulation (EU) 2026/1744 (Digital Omnibus on AI), OJ L 24 Jul 2026, in force 27 Jul 2026, EUR-Lex. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj>
18. Regulation (EU) 2016/679 (General Data Protection Regulation), EUR-Lex. <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

All web sources accessed and checked against PeritumAI's archived copies on 25 October 2026.