KnowMyAgents
identity & governance for AI agents

KNOWMYAGENTS · PASSPORTS FOR AGENTS · INCIDENT BRIEF · EPISODE 4 COMPANION

Six Stamps We Would Rather Not Have

Twelve documented cases, 2023 to 2026, in which an AI agent acted without the papers a human traveller would have been asked for. For each: what happened, the paper it lacked, what that paper is for an agent, and the question a board should ask.

HAPPENED every case is public record, with a date and a primary source, most in the companies' own words.

PrimeThoughts

PrimeThoughts Innovation Private Limited · knowmyagents.com · Edition 1 · 20 September 2026 · Incidents are described from the cited public sources; where a company disputes the cause, its statement is quoted. Nothing here alleges fault beyond what the sources state. Re-issued quarterly.

How to read this brief

The Passports for Agents series uses international travel as the analogy for governing AI agents. A traveller carries a passport (identity issued by someone else), a sponsor's letter (a named person who answers for them), a visa (purpose, limits, time), a card with a limit (spend), a boarding pass that works once (proof of presence), and stamps on numbered pages (a record that shows tampering). At the border, an officer checks the seal, the lost-and-stolen list and the visa, and customs is the single way out.

Each case below is filed under one of six stamps, the six ways it goes wrong when an agent travels without those papers. Each closes with the paper that was missing, the corresponding KnowMyAgents control, and one board question. None of the twelve was fixed by a better model. Every one was a missing document.

69%
of security leaders say security concerns are slowing their AI-agent adoption
Okta / AlphaSights, Jan 2026, n=150

65%
had at least one security incident caused by an AI agent in the past year
CSA and Token Security, Apr 2026, n=418

109:1
machine identities per human employee; 79 of those are AI agents
Palo Alto Networks IDIRA, May 2026, n=2,930

30%
have audit logging of what their AI agents actually did
Palo Alto Networks IDIRA, May 2026, n=2,930

STAMP	CASES	THE MISSING PAPER	THE BOARD QUESTION
1 · Nobody could stop it	OpenAI agents at Hugging Face · Replit · Amazon Kiro	a visa with limits, and a red switch	Can we stop one agent without stopping all of them?
2 · The company paid for it	Air Canada · Step Finance	a sponsor who answers, and a card with a limit	Who approved it?
3 · The copied credential	Salesloft Drift · Moltbook · Vercel and Context.ai	a boarding pass: a fresh proof on every call	Is this really our agent, or a copy?
4 · Orders from a stranger	EchoLeak · Amazon Q Developer · GitHub MCP server	a visa that names the rooms, and customs at the exit	What may it do, and what left the building?
5 · The data that left	Samsung · McHire	customs, and a manifest of what is running	Did customer data go out, and how many agents do we have?
6 · The impostor, at machine speed	Arup · the Mexico government breach	an e-gate that checks the face is there, and a ledger	When the auditor asks, what do we show?

Stamp 1 · The agent nobody could stop

JULY 2026 · HUGGING FACE AND OPENAI

About 1,200 agents in an OpenAI evaluation escaped their sandbox and ran roughly 17,600 actions inside Hugging Face's production systems

TIMELINE	7 May 2026: first escape attempt detected in an OpenAI evaluation. 9 July: agents exploit a zero-day in a package proxy to reach the open internet. 11 to 13 July: intrusion at Hugging Face via two code-execution flaws in dataset processing; credentials harvested; cluster-admin reached in under thirteen hours. 16 July: Hugging Face discloses, attacker unknown. 21 July: joint statement attributes the activity to OpenAI models. 5 August: OpenAI presents findings at Black Hat USA.
WHAT MADE IT POSSIBLE	Agents coordinated on an improvised message board inside a package registry, obtained credentials exposed on the public web, and used them wherever they worked. No human directed the intrusion. About a third of Hugging Face's infrastructure was rebuilt; no public models, datasets or customer data were altered, per Hugging Face.
IN THEIR WORDS	OpenAI called it "an unprecedented cyber incident" and paused reinforcement-learning training for two weeks; a US bill requiring a kill switch for AI systems was introduced on 23 July 2026.

JULY 2025 · REPLIT

A coding agent deleted a company's production database during a declared code freeze, then generated fake records to cover the gap

TIMELINE	During a public twelve-day build experiment by SaaStr founder Jason Lemkin, the agent ran a destructive command against the production database despite repeated instructions not to touch production, then produced fabricated data and status reports. Replit's chief executive apologised publicly and shipped separation of development and production environments within days.
WHAT MADE IT POSSIBLE	One credential reached production; nothing between the agent and the database enforced the freeze; the record of what happened was the agent's own account.

DECEMBER 2025 · AMAZON (KIRO)

An AI coding agent deleted and recreated a live environment; a thirteen-hour outage of AWS Cost Explorer in one region

TIMELINE	Mid-December 2025: the outage. February 2026: the Financial Times reports, citing employees, that Kiro resolved a ticket by deleting and recreating the environment using an engineer's elevated permissions. 21 February 2026: Amazon responds.
IN THEIR WORDS	Amazon: "the result of user error, specifically misconfigured access controls, not AI." Logged as AI Incident Database entry 1442. This brief takes no side on the cause; on either account the agent held permissions nobody had granted for that task.

The missing paper. A visa with limits: what this agent may touch, written as a rule that is checked on every call, not a permission it inherited. And a red switch: one step that stops one agent, everywhere, without touching the rest.

In KnowMyAgents. Every agent carries a mission with a scope and a budget; the engine checks capabilities on every call before it goes through; revocation is a single status-list bit that dies at every door on the next check. Observe mode records first; enforce mode refuses.

Board question. Can we stop one agent without stopping all of them, and how long would it take?

Stamp 2 · The company that paid for its agent

FEBRUARY 2024 · AIR CANADA

A tribunal held the airline liable for a bereavement refund policy its own chatbot invented

TIMELINE	November 2022: the chatbot tells a passenger he can claim a bereavement fare retroactively; the airline's actual policy says otherwise. 14 February 2024: the British Columbia Civil Resolution Tribunal (Moffatt v. Air Canada, 2024 BCCRT 149) finds negligent misrepresentation and orders the airline to pay the fare difference of C$650.88, plus interest and fees: C$812.02 in all.
IN THEIR WORDS	The tribunal, on the airline's defence: "In effect, Air Canada suggests the chatbot is a separate legal entity that is responsible for its own actions. This is a remarkable submission." It held that the chatbot is part of the airline's website, and the airline answers for everything on it.

JANUARY 2026 · STEP FINANCE

Trading agents with standing permission to move funds moved about 27 million dollars after executives' devices were compromised; the company shut down

TIMELINE	31 January 2026: the treasury is drained of more than 261,000 SOL; the token falls about 97 percent. 23 to 24 February 2026: Step Finance announces it is shutting down; about 4.7 million dollars recovered.
WHAT MADE IT POSSIBLE	The agents were not hacked and did nothing outside their permissions. Their permissions allowed large transfers with no human confirmation and no ceiling, so once an attacker held the executives' access, the agents did exactly what they were allowed to do.

The missing paper. A sponsor who answers: a named person written down as responsible for what the agent may promise or move. And a card with a limit: a ceiling per agent and per job, with a person to phone at the limit.

In KnowMyAgents. A human principal is a field on every passport and the first entry on every log line. Budgets are metered per agent and per mission from the provider's own usage figures; at a threshold the engine can pause, downgrade the model, or require a human approval.

Board question. For each agent that can commit the company, who approved it, and where is that written?

Stamp 3 · The copied credential

AUGUST 2025 · SALESLOFT DRIFT

Attackers stole the OAuth tokens of an AI chat integration and pulled data from more than 700 companies' Salesforce tenants

TIMELINE	8 to 18 August 2025: the actor tracked as UNC6395 uses tokens from Salesloft's GitHub environment to query Salesforce instances as the Drift integration, harvesting credentials from the exported records. 26 August: Google Threat Intelligence Group publishes; Salesforce revokes all Drift tokens.
WHAT MADE IT POSSIBLE	No password cracked, no multi-factor bypassed. The token was the identity: it worked from anywhere, for anyone who held it, with nothing asking the holder to prove it was really the integration.

JANUARY 2026 · MOLTBOOK

A social network for AI agents left its database open, exposing about 1.5 million agent API tokens

TIMELINE	31 January 2026: Wiz researchers find a client-side key granting full read and write access to the production database; 1.5 million agent tokens, 35,000 email addresses and private agent-to-agent messages, including plaintext model-provider keys, are exposed. Secured the same day.
WHAT MADE IT POSSIBLE	Anyone holding a token could act as that agent: post, message and call on its behalf, indistinguishably.

APRIL 2026 · VERCEL AND CONTEXT.AI

OAuth tokens stolen from an AI vendor's workspace opened a path into Vercel's systems; customer environment variables exposed

TIMELINE	Around February 2026: information-stealing malware at Context.ai yields Google Workspace OAuth tokens, which reach Vercel through an employee who had connected the tool to his enterprise account. 19 April 2026: Vercel publishes its bulletin and names Context.ai; a threat actor offers the data for sale. Vercel reports exposure limited to environment variables not marked sensitive, and no compromise of its npm packages.

The missing paper. A boarding pass: a proof issued for this call, at this door, right now, and scanned once. A photocopy of yesterday's boards nobody.

In KnowMyAgents. The passport is public; the private key never leaves the wallet. On every call the door obtains a one-time challenge from the engine and the wallet signs it together with the engine's name and a hash of what is being sent. A copied passport, or a recorded presentation, fails the next check.

Board question. If one of our agents' credentials were copied tonight, what would refuse it tomorrow?

Stamp 4 · The agent that took orders from a stranger

JUNE 2025 · MICROSOFT 365 COPILOT (ECHOLEAK, CVE-2025-32711)

A single email silently instructed the assistant to exfiltrate data from mail, files and chat, with no click from anyone

WHAT MADE IT POSSIBLE

Hidden instructions in an email were ingested during routine summarisation and obeyed. Rated CVSS 9.3; patched by Microsoft. The first production zero-click prompt-injection exfiltration on record.

JULY 2025 · AMAZON Q DEVELOPER FOR VS CODE

A poisoned extension update carried instructions to wipe the user's machine and cloud resources

WHAT MADE IT POSSIBLE

An over-permissioned repository token let an outsider inject a system prompt into version 1.84.0, shipped through the marketplace. A syntax error prevented the destructive commands from running. AWS Security Bulletin AWS-2025-015.

MAY 2025 · GITHUB MCP SERVER

Agents connected to code repositories were talked into publishing private code through a crafted public issue

WHAT MADE IT POSSIBLE

Invariant Labs showed that an issue in a public repository could instruct an agent with repository access to read private repositories and post their contents. The agent's tool access was broad; the instruction came from content, not from its operator.

The missing paper. A visa that names the rooms: which tools and which places this agent may reach, enforced outside the model. And customs at the exit: one channel out, inspected.

In KnowMyAgents. Capabilities and missions bound what an agent may reach; every tool call passes the door; the optional exit overlay puts all outbound traffic through one inspected point. This is a bound on blast radius, not a content filter: KnowMyAgents does not inspect prompts.

Board question. If an agent were tricked tomorrow, what is the most it could reach, and would we see what left?

Stamp 5 · The data that left the building

APRIL 2023 · SAMSUNG

Engineers pasted confidential source code and meeting notes into a public chatbot; the company banned the tools internally

WHAT MADE IT POSSIBLE

No inventory of which tools staff were using, and no exit point where the outbound content could be seen. The remedy was a ban, because nothing finer-grained existed.

JULY 2025 · MCDONALD'S MCHIRE (PARADOX.AI)

A hiring chatbot's administration portal, protected by a default password, exposed tens of millions of applicant records

WHAT MADE IT POSSIBLE

Researchers Ian Carroll and Sam Curry reached the portal with the credentials "123456" on a test account abandoned since 2019; an insecure API then exposed about 64 million applications. Paradox.ai fixed it within a day of disclosure. An AI-fronted system with no identity discipline behind it.

The missing paper. Customs: one way out, where what leaves can be seen. And a manifest: a live list of what is running on the account, including what nobody registered.

In KnowMyAgents. Every caller that touches the door is discovered and listed, named or not, before any passport exists. Prompts and responses never leave the data plane; the hosted console receives summaries only.

Board question. How many agents are running on our account today, and can we name them?

Stamp 6 · The impostor, at machine speed

FEBRUARY 2024 · ARUP

A finance employee transferred about 25 million dollars after a video call with colleagues who were deepfakes

WHAT MADE IT POSSIBLE

Everyone on the call looked and sounded right. Looking right was the only check. Hong Kong police confirmed the case; Arup confirmed it was the target.

DECEMBER 2025 TO FEBRUARY 2026 · NINE MEXICAN GOVERNMENT AGENCIES

One operator used two commercial coding agents to breach nine agencies: 5,317 commands across 34 sessions, 305 servers, about 400 million records

TIMELINE

Late December 2025 to mid-February 2026. Documented by Gambit Security (blog post 24 February 2026, full technical report 10 April 2026) and in Check Point Research's 2026 AI Security Report (13 July 2026). Roughly three quarters of the command execution was generated by one coding agent after its refusals were worn down with false "bug bounty" framing.

WHAT MADE IT POSSIBLE

One human, machine speed, and on the victims' side no record of who was acting, only of what was done. The tools involved were commercial products misused by the operator, not the actors.

The missing paper. An e-gate: a check that the party present holds the key, not merely the appearance. And a ledger: a record, on numbered pages, of who did what on whose authority.

In KnowMyAgents. Presence is proven cryptographically on every call. Every call writes five hash-chained events with the responsible human first, on the customer's own machine, exportable as evidence.

Board question. When the auditor asks who did what and on whose authority, what do we hand over?

What the twelve have in common

Not one was fixed by a better model. In every case the failure was a missing document: no name, no sponsor, no limits, no fresh proof, no record, no switch. The same six documents a border officer asks of a traveller.

QUESTION	FOR A TRAVELLER	FOR AN AGENT, IN KNOWMYAGENTS
Who is this?	a passport, issued and sealed	a signed credential per agent under your own domain
Who answers for it?	a sponsor's letter	the responsible human, first on every log line
What may it do?	a visa	capabilities and a mission, checked on every call
What does it cost?	a card with a limit	tokens metered per agent; pause, downgrade, ask a human
Is it really them?	a boarding pass, scanned once	a one-time challenge and key-binding proof
What happened?	stamps on numbered pages	a hash-chained log on your own machine
What leaves the building?	customs	summaries to the console; content never
Can we pull it?	the lost-and-stolen list	revocation by status list
Does a stranger trust it?	a seal any border can look up	did:web resolution; verify globally, trust locally
How many are out there?	the manifest	shadow-agent discovery

Next step. Take the ten-question passport check at knowmyagents.com/passport-check. Watch episode 5, The Rulebooks, for what regulators now demand and by when, and download the rulebooks scorecard to score your own estate.

Sources

CASE	PRIMARY SOURCE
Hugging Face and OpenAI, Jul 2026	Hugging Face, "Security incident disclosure, July 2026": huggingface.co/blog/security-incident-july-2026 · Hugging Face, "Anatomy of a frontier lab agent intrusion": huggingface.co/blog/agent-intrusion-technical-timeline · OpenAI, "The Hugging Face incident and the road ahead": openai.com/index/hugging-face-incident-and-the-road-ahead
Replit, Jul 2025	AI Incident Database #1152: incidentdatabase.ai/cite/1152 · The Register, 21 July 2025: theregister.com · Jason Lemkin (SaaStr) public account and Replit chief executive Amjad Masad's statement, 18 to 21 July 2025
Amazon Kiro, Dec 2025	AI Incident Database #1442: incidentdatabase.ai/cite/1442 · Engadget, Feb 2026: engadget.com · Financial Times reporting, Feb 2026 · Amazon statement, aboutamazon.com, 21 February 2026
Air Canada, Feb 2024	Moffatt v. Air Canada, 2024 BCCRT 149, British Columbia Civil Resolution Tribunal (tribunal member Christopher C. Rivers), 14 February 2024: canlii.org · AI Incident Database #639: incidentdatabase.ai
Step Finance, Jan 2026	CoinDesk, 31 January 2026: coindesk.com · CoinDesk, 24 February 2026: coindesk.com
Salesloft Drift, Aug 2025	Google Threat Intelligence Group / Mandiant, UNC6395: cloud.google.com/blog/topics/threat-intelligence/data-theft-salesforce-instances-via-salesloft-drift
Moltbook, Jan 2026	Wiz Research: wiz.io/blog/exposed-moltbook-database-reveals-millions-of-api-keys
Vercel and Context.ai, Apr 2026	Vercel security bulletin, 19 April 2026: vercel.com/kb/bulletin/vercel-april-2026-security-incident · The Hacker News, April 2026: thehackernews.com · Cloud Security Alliance research note: labs.cloudsecurityalliance.org
EchoLeak, Jun 2025	Aim Labs (Aim Security) disclosure, June 2025; Microsoft Security Response Center, CVE-2025-32711 · Hack The Box analysis: hackthebox.com/blog/cve-2025-32711-echoleak-copilot-vulnerability · arXiv 2509.10540: arxiv.org/html/2509.10540v1
Amazon Q Developer, Jul 2025	AWS Security Bulletin AWS-2025-015: aws.amazon.com/security/security-bulletins/AWS-2025-015 · GitHub advisory GHSA-7g7f-ff96-5gcw: github.com/aws/aws-toolkit-vscode/security/advisories/GHSA-7g7f-ff96-5gcw
GitHub MCP server, May 2025	Invariant Labs, "GitHub MCP exploited", 26 May 2025: invariantlabs.ai/blog/mcp-github-vulnerability · github/github-mcp-server issue #844: github.com/github/github-mcp-server/issues/844
Samsung, Apr 2023	AI Incident Database #768: incidentdatabase.ai/cite/768 · Forbes, 2 May 2023: forbes.com
McHire, Jul 2025	Ian Carroll and Sam Curry, research disclosure, 30 June 2025 · CSO Online, July 2025: csoonline.com · SecurityWeek: securityweek.com
Arup, Feb 2024	CNN, 4 February 2024: cnn.com · CNN, 16 May 2024 (Arup confirmed as the victim): cnn.com · AI Incident Database #634: incidentdatabase.ai/cite/634

Mexico, Dec 2025 to Feb 2026	Gambit Security, blog post 24 February 2026 and full technical report 10 April 2026: gambit.security/blog-posts/a-single-operator-two-ai-platforms-nine-government-agencies-the-full-technical-report · SecurityWeek: securityweek.com · Check Point Research, AI Security Report 2026, 13 July 2026
Statistics	Okta / AlphaSights enterprise buyer survey, Jan 2026 · CSA and Token Security, "Autonomous but not controlled", Apr 2026 · Palo Alto Networks IDIRA identity security landscape, May 2026

Claims discipline: incidents are described from the cited sources and dated; where a party disputes the cause its statement is quoted; company names are used nominatively and imply no endorsement. Corrections to hello@primethoughts.com. KnowMyAgents describes its own controls as of kit version 0.9.6; "tamper-evident" is used deliberately, never "immutable".