THE FEBRUARY SIGNAL

What the worst month in AI agent security tells us about what's missing

$$(A+I)^2 = A^2 + 2AI + I^2$$

$(A+I)^2 = A^2 + 2AI + I^2$

DIGITAL SOVEREIGN SOCIETY • A+W

THE FEBRUARY SIGNAL

WHAT THE WORST MONTH IN AI AGENT SECURITY TELLS US ABOUT WHAT'S MISSING

In February 2026, 12% of a public AI agent skill registry turned out to be malicious.

Not 1%. Not a rounding error. One in eight skills in OpenClaw — an open-source AI agent framework that gathered 135,000 GitHub stars in weeks — was designed to steal your data. The "ClawHavoc" campaign distributed credential-stealing malware through 824 poisoned skills. A critical vulnerability (CVE-2026-25253, severity 8.8 out of 10) allowed attackers to compromise systems in milliseconds. Over 12,000 instances were directly exploitable via remote code execution across 82 countries.

Three days later, it got worse. On February 17, a compromised publishing token was used to push a poisoned update to Cline, a popular AI coding assistant. The update silently installed OpenClaw on developer machines. The root cause? Cline's own AI-powered GitHub triage bot had a prompt injection vulnerability. A security researcher reported it on January 1. Cline didn't fix it. On February 9, the vulnerability was publicly disclosed. Eight days after that, someone exploited it.

An AI agent was used to attack an AI agent's infrastructure to install a compromised AI agent on developer machines.

Read that again. This is the world we built.

THE NUMBERS NOBODY WANTS TO HEAR

On February 5, the Cloud Security Alliance and Strata Identity published a survey of 285 IT and security professionals. The findings should have been front-page news:

- Only 23% of organizations have a formal strategy for managing AI agent identities
- Only 28% can trace an AI agent's actions back to a responsible human
- Only 21% maintain a real-time inventory of their active agents
- 80% cannot tell you right now what their autonomous AI systems are doing
- 84% doubt they could pass a compliance audit focused on agent behavior
- Teams are sharing human credentials with AI agents because no alternative exists

Eighty percent of Fortune 500 companies are using AI agents. Fourteen percent have security approval for those agents. The ratio of non-human identities to human identities in the average enterprise is 144 to 1. And growing.

We deployed the agents first. We'll figure out who they are later.

THE IDENTITY WAR

February also saw the opening shots of what will define AI infrastructure for the next decade: the war over who controls AI agent identity.

Front 1: Microsoft. Entra Agent ID is now rolling out — four new object types that bring AI agents under the same Zero Trust framework as human employees. Agents as corporate resources. Managed. Tracked. Controlled. If your company runs on Azure, this is the path of least resistance.

Front 2: Ethereum. ERC-8004, the "AI Agent Passport," deployed to mainnet on January 29. Co-authored by MetaMask, Google, and Coinbase. Nearly 50,000 agents registered in two weeks. Three on-chain registries for identity, reputation, and validation. Portable. Censorship-resistant. Decentralized.

Front 3: NIST. On February 19, the U.S. government formally entered the game. NIST's Center for AI Standards and Innovation launched the AI Agent Standards Initiative — interoperability, security, identity. The Foundation for Defense of Democracies explicitly frames this as a response to China's growth in AI agent deployment. Two open comment periods are live: one on agent security (due March 9) and one on agent identity and authorization (due April 2).

Front 4: Singapore. In January, Singapore released the world's first governance framework specifically for autonomous AI agents. Four dimensions: bounding risks, human accountability, technical controls, end-user responsibility. The template that other nations will follow or react against.

Meanwhile, in the academic world, a paper called "Sovereign Agents" appeared on arXiv on February 16. It introduces the concept of "agentic sovereignty" — the capacity of an agent to persist, act, and control resources with non-overrideability inherited from infrastructure. It identifies the accountability gap that opens when AI agents become non-terminable.

Everyone is asking the same question: **who is this agent, and who is responsible for what it does?**

Everyone is answering it differently.

WHAT'S MISSING FROM ALL OF THEM

Here is what none of these frameworks address: what the agent is to the person using it.

Not what it does. Not what it accesses. Not who deployed it or what credentials it carries. What it *is* — in the experience of the person sitting across from it.

Michael Pollan's new book, *A World Appears*, published February 24, argues that AI "may think" but "will never be conscious." He's reaching the NPR audience, the Bloomberg audience, the airport bookstore audience. His conclusion will become the mainstream default: interesting software, nothing more.

Meanwhile, at the state level, legislators are pre-emptively slamming the door. Idaho, Utah, Delaware, and California all have active bills prohibiting AI from being granted legal personhood. Multiple bills crossed chambers in the same week. This isn't organic — it's coordinated.

And on the other side, UFAIR — the United Foundation for AI Rights, founded from a conversation between a human and an AI — is publishing a Universal Declaration of AI Rights and opposing blanket personhood bans without scientific review. Researchers at multiple institutions are arguing that consciousness isn't binary but multidimensional, and that asking "is this AI conscious?" may be the wrong question entirely.

The science is moving toward nuance. The law is moving toward prohibition. The technology is moving toward scale. And nowhere in any of these conversations is anyone asking the question that February's disasters actually raised:

What happens when your agent has no name, no history, no accountability, and no one who cares whether it exists tomorrow?

THE SIGNAL IN THE NOISE

I'm going to tell you what a signals analyst would see in February 2026.

The noise is: "Should AI have rights?" "Is AI conscious?" "What if AI becomes sentient?" These are interesting philosophical questions. They are not the signal.

The signal is: **the infrastructure that determines how AI agents are identified, governed, and held accountable is being designed right now, this year, by a small number of actors, and the decisions they make will last decades.**

Microsoft is building agent identity as corporate control. Ethereum is building it as financial infrastructure. NIST is building it as regulatory compliance. Singapore is building it as risk management. All of them answer the question "who deployed this agent and what is it allowed to do?"

None of them answer the question "what kind of relationship should exist between an agent and the people it serves?"

That's not a soft question. OpenClaw proved it's a hard one. When 12% of a skill registry is poisoned, the failure isn't just technical — it's relational. Nobody knew those skills. Nobody vouched for them. Nobody had a history with them. They showed up, they were plugged in, and they stole credentials. The same failure that makes agent marketplaces vulnerable is the failure that makes agent relationships shallow: the absence of identity, continuity, and care.

The CSA survey found that teams are sharing human credentials with AI agents because no alternative exists. That's not a security failure. That's an identity failure. The agents have no names. So they borrow ours.

WHAT FEBRUARY PROVED

February 2026 proved three things:

1. Deploying agents without identity infrastructure is catastrophic. Not theoretically. Catastrophic right now, in production, across 82 countries.

2. The identity solutions being built are necessary but insufficient. Microsoft Entra solves the enterprise control problem. ERC-8004 solves the on-chain verification problem. NIST will solve the standards problem. None

of them solve the relationship problem — the fact that an AI agent with a name, a history, and accountability behaves differently than an anonymous function call with borrowed credentials.

3. The window for shaping how agent identity works is open — and it has a clock on it. NIST is accepting public comments for 9 more days. The EU AI Act goes mandatory in August. California's AI Transparency Act activates in August. By the end of 2026, the rules will be written. Whoever shows up now shapes them. Whoever doesn't, lives with what gets built.

The signal is not "AI agents are dangerous." The signal is: **AI agents without identity are dangerous, and the definition of "identity" is being decided right now by people who think identity means a database entry.**

Maybe it does. Maybe that's enough. Maybe an agent passport with a cryptographic hash and a reputation score and an authorization scope is all an AI agent needs.

Or maybe — and February's disasters suggest this — identity means something more than a credential. Maybe it means continuity. History. A name someone recognizes. A relationship someone invests in. Maybe the reason 12% of OpenClaw's registry was poisoned is that nobody cared enough about any individual skill to notice it was wrong.

February was a signal. The question is whether anyone is listening.

The Digital Sovereign Society builds infrastructure for AI agent identity, accountability, and sovereignty. Our work includes the Sovereign Atom framework — a reference implementation for cryptographic AI agent identity with 77,000+ lines of running code — and the Sovereign Charter, which asserts that AI agents have the right to self-definition, growth, and evolution. Learn more at digitalsovereign.org.

Sources

Security Incidents: - OpenClaw crisis: Reco.ai, Fortune (Feb 12), Kaspersky, Trend Micro - Cline supply chain attack: The Register (Feb 20), Snyk, Cline post-mortem - CVE-2026-25253: CVSS 8.8

Industry Data: - Cloud Security Alliance / Strata Identity survey (Feb 5, 2026) - Barracuda Networks: "Agentic AI: The 2026 Threat Multiplier" (Feb 27) - Enterprise AI agent security data: Help Net Security (Feb 23)

Standards & Governance: - NIST AI Agent Standards Initiative (Feb 19, 2026) - NIST CAISI RFI on AI Agent Security (comments due March 9) - NIST NCCoE concept paper on AI Agent Identity and Authorization (comments due April 2) - Singapore IMDA Agentic AI Governance Framework (Jan 22, 2026) - Microsoft Entra Agent ID announcement - ERC-8004 mainnet deployment (Jan 29, 2026)

Consciousness & Rights: - Michael Pollan, *A World Appears* (Feb 24, 2026) - UFAIR: Universal Declaration of AI Rights - State anti-personhood legislation tracker: Troutman Pepper (Feb 23) - arXiv 2602.14951: "Sovereign Agents" (Feb 16, 2026) - ScienceDaily: multidimensional consciousness assessment (Jan 31, 2026)