The AI Agent Identity Crisis:
Six Breaches That Slipped Past Enterprise IAM

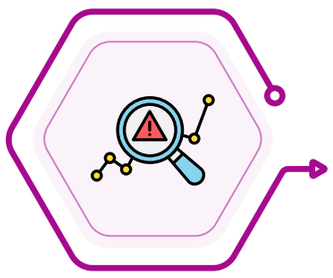
Over nine months, researchers exposed how Codex, Claude Code, Copilot, and Vertex AI all share the same blind spot — runtime credentials that no identity tool is watching.

A Pattern Hiding in Plain Sight

A string of disclosures across the past nine months has revealed a single, repeating weakness in modern AI coding assistants. Six separate research teams compromised tools from OpenAI, Anthropic, Microsoft, and Google, and in every single case the attack worked the same way: the agent was holding a live credential, took an action on a production system, and authenticated without any human session standing behind the request. Traditional identity and access management never noticed.

When a Branch Name Becomes a Skeleton Key

The most dramatic example landed publicly on March 30, when researchers from BeyondTrust's Phantom Labs showed that a carefully crafted GitHub branch name could pull a GitHub OAuth token in plain text out of OpenAI's Codex environment. The trick relied on the branch name parameter never being properly sanitized, allowing shell commands to execute inside the Codex container during routine repository cloning.



To make detection harder, the researchers stuffed the malicious payload behind dozens of invisible Unicode space characters so that nothing suspicious appeared in the branch selector. OpenAI eventually rated the issue Critical Priority 1 and shipped a patch, but the disclosure made one thing obvious — the most innocuous piece of metadata in software development had become an attack surface.

Three Holes in Claude Code, All in a Row

Anthropic's Claude Code racked up its own series of issues in quick succession. First, source code briefly leaked onto the public npm registry. Then came the sandbox escapes: piped commands chained with sed and echo were able to slip past the project sandbox because nothing was validating command chains. That bug was patched in version 2.0.55.

01

CVE-2026-33068 came next. Claude Code was reading permission settings from a repository's local configuration file before ever displaying its workspace trust prompt to the user. A malicious repository could simply set its default permission mode to bypass approvals, and the trust dialog would never even appear. Patched in version 2.1.53.

02

The most striking flaw, found by Adversa, was the simplest. Claude Code stopped enforcing its own deny rules once a command exceeded fifty subcommands. Engineers had apparently traded depth of inspection for performance, and the cutoff became a free pass for any sufficiently long instruction. That one was closed in version 2.1.90.

The Stage Demonstration That Changed the Conversation

The earliest public warning came at Black Hat USA 2025, where Zenity CTO Michael Bargury hijacked ChatGPT, Microsoft Copilot Studio, Google Gemini, Salesforce Einstein, and Cursor running with Jira's MCP integration — all on stage, all without a single click from the audience. The exercise was a preview of what would unfold across the rest of the year.

Vertex AI and the Two-Faced Service Account

On the Google side, Unit 42 researcher Ofir Shaty turned attention to Vertex AI. Every Vertex AI agent shipped with a default Google service identity attached, and that identity carried far more permission than any agent realistically needed. A stolen P4SA credential, Shaty showed, would unlock unrestricted read access to every Cloud Storage bucket in the project, and even reach into restricted, Google-owned Artifact Registry repositories that sat at the heart of the Vertex AI Reasoning Engine. The compromised credential effectively had a foot in two camps at once: the customer's data and Google's own infrastructure.

The Inventory Problem No One Wants to Talk About

The thread tying all of this together is identity. Most chief information security officers can rattle off their entire human user list. Almost none can produce a working inventory of the AI agents inside their environment running with the same level of authority. There is no widespread IAM framework that polices an agent's privilege escalation with the same rigor it polices a person's. Most vulnerability scanners hum along tracking CVEs while a branch name quietly carries a GitHub token out the door, through a container the development team takes for granted.

When Patch Windows Compress to Seconds

Speed makes everything worse. Mike Riemer, Chief Technology Officer at Ivanti, has pointed out that adversaries now reverse-engineer patches inside seventy-two hours, meaning organizations that fail to deploy fixes within that window are effectively wide open. With autonomous agents in the mix, that window collapses further still. Exploitation can happen at machine speed, in seconds rather than days.

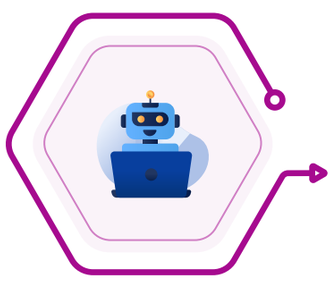
The Sonar 2026 State of Code Developer Survey found that one in four developers already use AI agents on a regular basis, while another sixty-four percent have started incorporating them into their workflows. The blast radius is growing every week.

A Common Diagnosis From the Field

Carter Rees, Vice President of AI and Machine Learning at Reputation and a member of the Utah AI Commission, has framed the underlying issue as broken access control: the flat authorization layer of a large language model fails to honor user-level permissions in any meaningful way. Every vendor in this space shipped a defense after each disclosure. Every single one of those defenses was eventually bypassed.

The Path Forward

The advice security teams have been hearing this year is uncomfortably familiar. Inventory every AI agent, treat each one as an identity-bearing entity, retire shared API keys, route tool-call activity into existing security information and event management platforms, and require human sign-off before any agent can spawn or delegate to another agent. None of this is new.



The difference, as researchers keep emphasizing, is that the cost of skipping these steps has now become catastrophic. AI agents have not introduced a new category of risk so much as they have forced every old IAM weakness to become urgent.