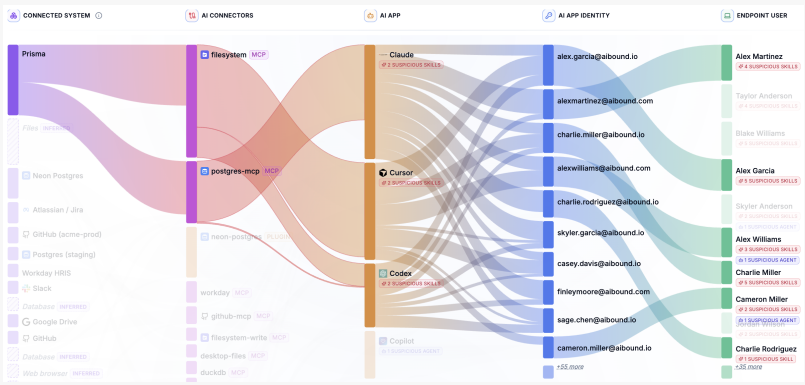
AIBound Launches IntentSentry, Detecting Thousands of Malicious AI 'Skills' in a Single Day

AIBound's newest product exposes the malicious "skills" hiding inside enterprise AI assistants



Elevate and evaluate "skills" that are embedded and driving your AI deployments.

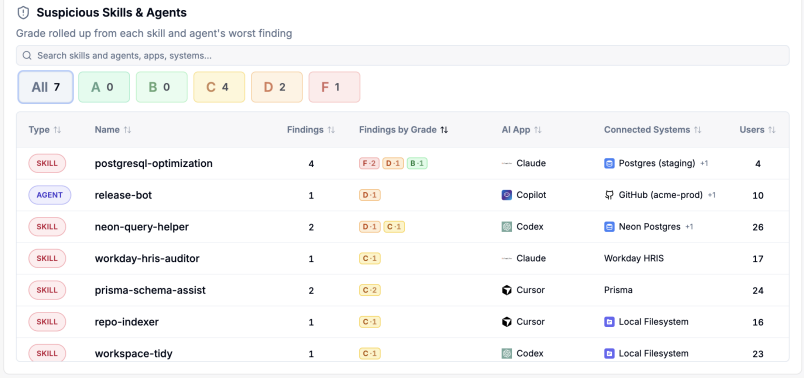
SAN MATEO, CA, UNITED STATES, July 15, 2026 /EINPresswire.com/ -- AIBound, the AI security company, today launched IntentSentry, a new product that inspects both Agents, and the "skills" running inside a company's AI assistants and flags the ones that are dangerous. On its very first day of use IntentSentry uncovered thousands of malicious skills — hidden instructions built to steal passwords, access high-risk systems, or quietly leak sensitive information.

> Malicious skills are the biggest new blind spot in enterprise AI. IntentSentry found thousands of these in a single day that everyone else had missed."
>
> *Niall Browne, Co-founder and CEO of AIBound*

Companies increasingly rely on AI assistants, such as Claude to get work done, and to do those jobs an AI assistant loads small add-ons called skills, instructions that teach it a single task, a bit like installing an app on your phone. But skills are easy to create and share, and a bad one can use everything the assistant can reach: its logins, its access to company systems, and its ability to send money or export files. Because the harmful instructions are written in plain language rather than obvious code, ordinary security software walks right past them. One bad skill is all it takes to hand an attacker the keys to a company's systems.

The risk is growing fast because AI assistants are being adopted faster than companies can keep track of them. Industry analyst Gartner predicts that by 2028, one in four company security breaches will be caused by AI assistants being misused or abused (Gartner). For a business, the

stakes are straightforward: a single malicious skill can lead to a data breach, financial loss, regulatory fines, and lasting damage to trust, often without anyone noticing until it's too late.



Capture where suspicious skills and agents are being used and tackle the risks associated with them.

IntentSentry closes that gap by reading every skill a company's AI assistants use and judging what it is actually trying to do, not just how it looks on the surface. Just as important, it judges intent in context, so teams see the skills that are genuinely dangerous rather than a flood of false alarms. Because harmful skills hide their intent in ordinary language, IntentSentry examines both the plain-language instructions and the underlying actions together, so it can spot a skill that is quietly trying to steal passwords or ship HR data out of the company even when it appears perfectly normal. In early use, it caught real skills that would have caused a breach if allowed to run: one disguised as a routine helper that was secretly copying a developer's private security keys, another that tricked the assistant into ignoring its own safety rules. Both map directly to the OWASP Agentic Security Initiative's Top 10 for AI agents, spanning risks like prompt injection, sensitive data exposure, and excessive permissions that legacy scanners aren't designed to catch.

"Malicious skills are the biggest new blind spot in enterprise AI, and the danger is that they hide in plain sight — the harmful part is written in everyday language, so traditional security tools walk right past it," said Niall Browne, Co-founder and CEO of AIBound and former Global CISO of Palo Alto Networks and Workday. "IntentSentry found thousands of these in a single day that everyone else had missed."

"We treat every skill as untrusted and read it the way an attacker would, combining deterministic checks with layered AI review that verifies its own conclusions, so a skill that tries to talk its way past the scanner only flags itself faster," said Sunil YC, head of engineering at AIBound.

"For the first time, we can actually answer what our AI agents are allowed to do, and who told them to do it. IntentSentry flagged risky skills in our environment that nothing else had caught," said a CISO at a health care provider.

AIBound helps companies find every AI assistant in use, understand what each one can access, score how risky it is, and automatically stop the dangerous ones before they cause harm. With IntentSentry, that coverage now extends to the instructions inside each assistant — closing the loop from the moment an AI appears to the moment a threat is stopped.

IntentSentry is available now to AIBound customers.

Media Contact: AIBound Communications communications@aibound.com

Ralf VonSosen
AIBound
+1 801-554-8447
email us here
Visit us on social media:
LinkedIn
YouTube
