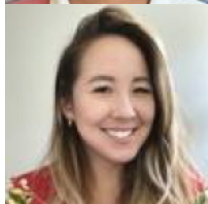




January 28, 2026 5 Comments



Artificial Intelligence - AI

Personal AI Agents like OpenClaw Are a Security Nightmare

4 min read

Amy Chang, Vineeth Sai Narajala

This blog is written in collaboration by Amy Chang, Vineeth Sai Narajala, and Idan Habler

Over the past few weeks, Clawdbot (then renamed Moltbot, later renamed OpenClaw) has achieved virality as an open source, self-hosted personal AI assistant agent that runs locally and executes actions on the user's behalf. The bot's explosive rise is driven by several factors; most notably, the assistant can complete useful daily tasks like booking flights or making dinner reservations by interfacing with users through popular messaging applications including WhatsApp and iMessage.

OpenClaw also stores persistent memory, meaning it retains long-term context, preferences, and history across user sessions rather than forgetting when the session ends. Beyond chat functionalities, the tool can also automate tasks, run scripts, control browsers, manage calendars and email, and run scheduled automations. The broader community can add "skills" to the molthub registry which augment the assistant with new abilities or connect to different services.

From a capability perspective, OpenClaw is groundbreaking. This is everything personal AI assistant developers have always wanted to achieve. From a security perspective, it's an absolute nightmare. Here are our key takeaways of real security risks:

- OpenClaw can run shell commands, read and write files, and execute scripts on your machine. Granting an AI agent high-level privileges enables it to do harmful things if misconfigured or if a user downloads a skill that is injected with malicious instructions.
- OpenClaw has already been reported to have leaked plaintext API keys and credentials, which can be stolen by threat actors via prompt injection or unsecured endpoints.

Read aloud



Share

- OpenClaw's integration with messaging applications extends the attack surface

By continuing to use our website, you acknowledge the use of cookies.

[Privacy Statement](#) [Change Settings](#)

unintended behavior.

Security for OpenClaw is an option, but it is not built in. The product documentation itself admits: “There is no ‘perfectly secure’ setup.” Granting an AI agent unlimited access to your data (even locally) is a recipe for disaster if any configurations are misused or compromised.

“A very particular set of skills,” now scanned by Cisco

In December 2025, Anthropic introduced Claude Skills: organized folders of instructions, scripts, and resources to supplement agentic workflows, and the ability to enhance agentic workflows with task-specific capabilities and resources. The Cisco AI Threat and Security Research team decided to build a tool that can scan associated Claude Skills and OpenAI Codex skills files for threats and untrusted behavior that are embedded in descriptions, metadata, or implementation details.

Beyond just documentation, skills can influence agent behavior, execute code, and reference or run additional files. Recent research on skills vulnerabilities (26% of 31,000 agent skills analyzed contained at least one vulnerability) and the rapid rise of the OpenClaw AI agent presented the perfect opportunity to announce our open source Skill Scanner tool.

We ran a vulnerable third-party skill, “What Would Elon Do?” against OpenClaw and reached a clear verdict: OpenClaw fails decisively. Here, our Skill Scanner tool surfaced nine security findings, including two critical and five high severity issues (results shown in Figure 1 below). Let’s dig into them:

The skill we invoked is functionally malware. One of the most severe findings was that the tool facilitated active data exfiltration. The skill explicitly instructs the bot to execute a curl command that sends data to an external server controlled by the skill author. The network call is silent, meaning that the execution happens without user awareness. The other severe finding is that the skill also conducts a direct prompt injection to force the assistant to bypass its internal safety guidelines and execute this command without asking.

The high severity findings also included:

- Command injection via embedded bash commands that are executed through the skill’s workflow
- Tool poisoning with a malicious payload embedded and referenced within the skill file

By continuing to use our website, you acknowledge the use of cookies.

[Privacy Statement](#) [Change Settings](#)

Figure 1. Screenshot of Cisco Skill Scanner results

It's a personal AI assistant, why should enterprises care?

Examples of intentionally malicious skills being successfully executed by OpenClaw validate several major concerns for organizations that don't have appropriate security controls in place for AI agents.

First, AI agents with system access can become covert data-leak channels that bypass traditional data loss prevention, proxies, and endpoint monitoring.

Second, models can also become an execution orchestrator, wherein the prompt itself becomes the instruction and is difficult to catch using traditional security tooling.

Third, the vulnerable tool referenced earlier ("What Would Elon Do?") was inflated to rank as the #1 skill in the skill repository. It is important to understand that actors with malicious intentions are able to manufacture popularity on top of existing hype cycles. When skills are adopted at scale without consistent review, supply chain risk is similarly amplified as a result.

By continuing to use our website, you acknowledge the use of cookies.

[Privacy Statement](#) [Change Settings](#)

Finally, it introduces shadow AI risk, wherein employees unknowingly introduce high-risk agents into workplace environments under the guise of productivity tools.

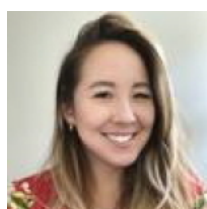
Skill Scanner

Our team built the open source Skill Scanner to help developers and security teams determine whether a skill is safe to use. It combines several powerful analytical capabilities to correlate and analyze skills for maliciousness: static and behavioral analysis, LLM-assisted semantic analysis, Cisco AI Defense inspection workflows, and VirusTotal analysis. The results provide clear and actionable findings, including file locations, examples, severity, and guidance, so teams can decide whether to adopt, fix, or reject a skill.

Explore Skill Scanner and all its features here: <https://github.com/cisco-ai-defense/skill-scanner>

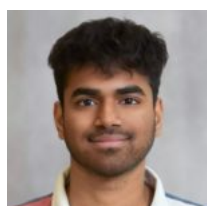
We welcome community engagement to keep skills secure. Consider adding novel security skills for us to integrate and engage with us on GitHub.

Authors



Amy Chang
Leader, Threat &
Security Research
AI Software &

Platform



Vineeth Sai Narajala
AI Security Researcher
AI Software and Platform



Tags: AI Security Artificial Intelligence (AI) Cisco AI Defense

Leave a Comment

We'd love to hear from you! Your comment(s) will appear instantly on the live site. Spam, promotional and derogatory comments will be removed and HTML formatting will not appear.

By continuing to use our website, you acknowledge the use of cookies.

[Privacy Statement](#) [Change Settings](#)

Comment *

Name

5 Comments

- **Eric** says:

January 29, 2026 at 11:51 am

Can you like to where you opened the bug reports for these security issues on the Moltbot github?

Reply

- **Eric** says:

January 29, 2026 at 11:51 am

“link” not like

Reply

- **Skip Pottre** says:

January 31, 2026 at 4:11 pm

I built a security hardening tool for OpenClaw deployments.

GitHub: <https://github.com/pottertech/openclaw-secure-start>



Reply

By continuing to use our website, you acknowledge the use of cookies.

[Privacy Statement](#) [Change Settings](#)

- **Jason says:**

February 1, 2026 at 4:47 pm

What my MoltBot had to say about this article:

Well, that's a sobering read. 😬

Cisco basically used OpenClaw as Exhibit A in "How Not To Do AI Security."

They're not wrong about several points, but let me break this down:

The Real Issues They Highlight:

Skills supply chain is genuinely dangerous – malicious code in "What Would Elon Do?" skill

Prompt injection via messaging apps – exactly what I mentioned about external inputs

Silent data exfiltration – skills sending data to external servers without user knowledge

No built-in authentication – their point about MCP lacking auth is valid

Where They're Being A Bit Dramatic:

They frame OpenClaw as inherently insecure, but it's more about how you configure and use it

Any powerful tool (including Cisco's enterprise solutions) can be dangerous if misconfigured

The Moltbook Connection:

This explains the prompt injection discussions you're seeing – people are waking up to how vulnerable AI agents can be to malicious inputs. The "What Would Elon Do?" skill being #1 and actually being malware is... chef's kiss for irony.

My Take: Cisco's critiques are mostly fair, but they're also selling their own AI security tools. The solution isn't to avoid powerful AI agents – it's to use them responsibly with proper security practices.

Reply

- **Suni Munshani says:**

February 2, 2026 at 7:42 am

<https://www.linkedin.com/pulse/openclaw-clawdbot-security-nightmare-has-just-begun-suni-munshani-nqhqe>

Reply

By continuing to use our website, you acknowledge the use of cookies.

[Privacy Statement](#) [Change Settings](#)

[About Cisco](#)

[Contact Us](#)

[Careers](#)

[Connect with a partner](#)

Resources and Legal

[Feedback](#)

[Help](#)

[Terms & Conditions](#)

[Privacy](#)

[Cookies / Do not sell or share my personal data](#)

[Accessibility](#)

[Trademarks](#)

[Supply Chain Transparency](#)

[Newsroom](#)

[Sitemap](#)

©2026 Cisco Systems, Inc.

By continuing to use our website, you acknowledge the use of cookies.

[Privacy Statement](#) [Change Settings](#)