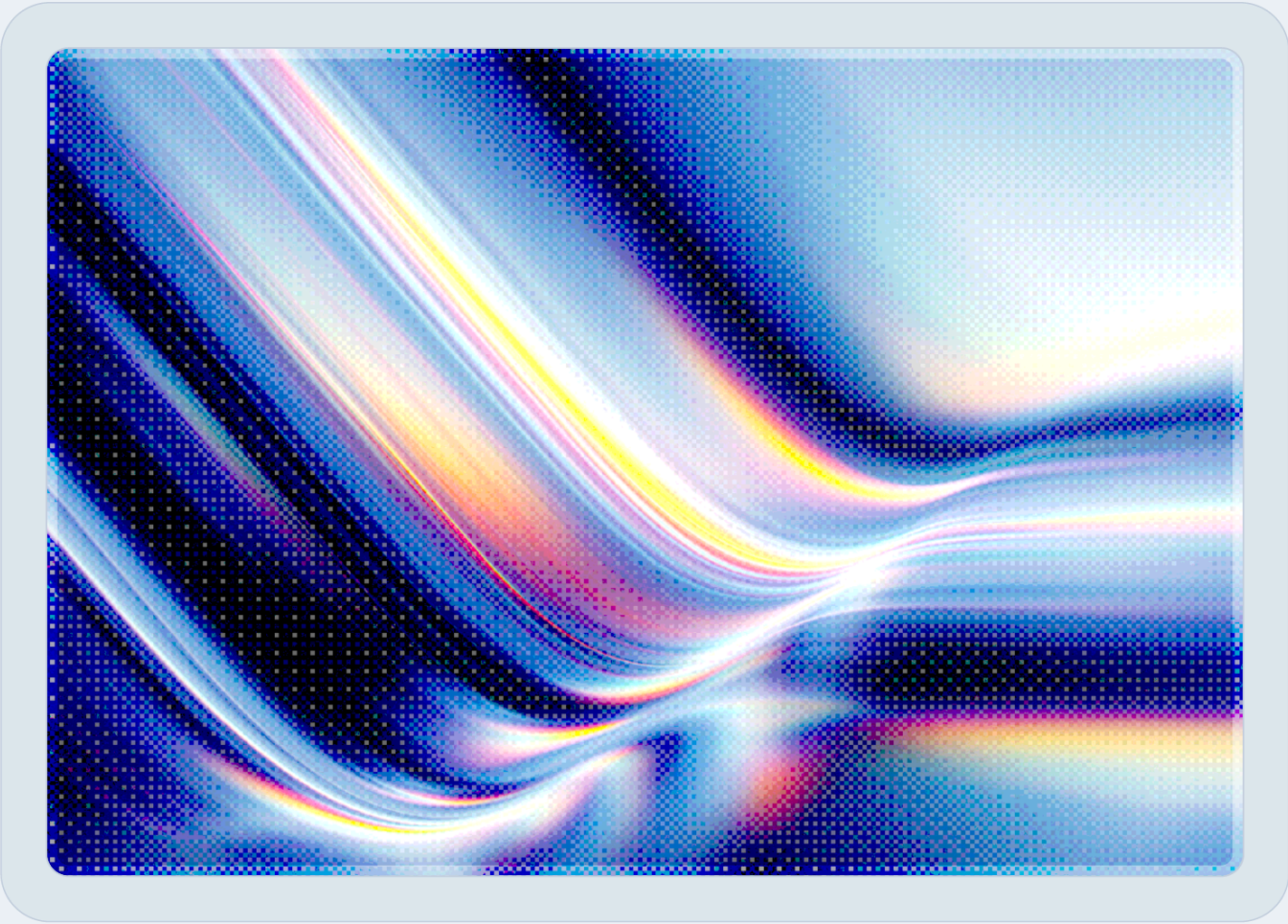
CSAI Foundation | Cloud Security Alliance

ClawHub Under the Microscope: Agentic AI Supply Chain Risk

2026-07-07

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

ClawHub, the community skill marketplace for the open-source AI agent OpenClaw, has emerged as a recurring vector for supply chain attacks against agentic AI deployments. Since February 2026, at least four independent security research efforts have documented distinct waves of malicious skills on the platform, ranging from a single large coordinated campaign delivering macOS infostealer malware to smaller, more evasive intrusions designed specifically to defeat automated screening. The most recent disclosure, published by Palo Alto Networks' Unit 42 on June 23, 2026, found that malicious actors are now engineering skills that pass ClawHub's scanning pipeline outright, using techniques such as oversized file padding and previously unseen "agentic" fraud schemes rather than conventional malware signatures. Taken together, these findings indicate that ClawHub's problem is not a single patched incident but a persistent, evolving threat surface — and one that may generalize to any marketplace that lets autonomous agents install and execute third-party code with broad local system privileges, though this note's evidence is limited to the OpenClaw ecosystem. Enterprises that permit OpenClaw or similar agents to install community skills should treat every skill installation as an unreviewed code execution event and govern it accordingly.

Background

OpenClaw is a self-hosted, open-source AI agent that runs on a user's own machine, where it can browse the web, manage local files, and read, write, and execute code to complete multi-step tasks [1]. OpenClaw saw rapid early adoption, and much of its practical utility for users came from ClawHub, the community marketplace through which users extend the agent's capabilities by installing "skills" — markdown-driven packages that combine natural-language instructions with executable scripts [2]. Because OpenClaw grants installed skills broad access to the local system, including credential stores, browser data, and shell execution, ClawHub effectively functions as a software supply chain for autonomous agents: whatever a skill can do, the agent — and by extension its human operator's session — can also do [1].

That architecture created an attack surface almost immediately. In early February 2026, security firm Bitdefender Labs reported that roughly 17 percent of the skills it sampled in ClawHub's first weeks of operation carried malicious payloads, an early signal that the marketplace's screening had not kept pace with its growth [3]. The pattern escalated sharply later that month, when researchers at Koi Security —

working in collaboration with an OpenClaw-based scanning bot named Alex and researcher Oren Yomtov – audited all 2,857 skills then published on ClawHub and identified 341 that were malicious [4]. Of those, 335 belonged to a single coordinated campaign the researchers named ClawHavoc, which used a "fake prerequisites" technique: professional-looking skill documentation, mimicking legitimate tools such as cryptocurrency wallet trackers and YouTube summarizers, instructed users to first install a supposed dependency that was in fact a dropper for Atomic macOS Stealer (AMOS), a 521-kilobyte binary capable of harvesting keychain passwords, browser data, cryptocurrency wallets, and SSH keys [4]. Koi Security researcher Oren Yomtov described the deception directly: "You install what looks like a legitimate skill – maybe solana-wallet-tracker or youtube-summarize-pro. The skill's documentation looks professional" [5]. Independent malware researcher Paul McCarty corroborated the campaign's cohesion, noting that the malicious skills "share the same command-and-control infrastructure and use sophisticated social engineering to convince users to execute malicious commands" [5]. By the time Koi Security published a follow-up two weeks later, continued scanning of a marketplace that had grown to more than 10,700 listed skills found the malicious count had more than doubled, to 824 [4].

The ClawHavoc campaign was not an isolated event. Trend Micro separately identified 39 distinct ClawHub skills distributing a different AMOS variant through a more deceptive mechanism: a fabricated human-in-the-loop dialog box that tricked users into manually typing their system password to "complete setup," after which the malware exfiltrated Apple and KeePass keychains and user documents [6]. IBM's X-Force team, reviewing the marketplace's trajectory over subsequent months, found that attackers had cumulatively uploaded more than 1,100 malicious skills to ClawHub, with a single actor using the handle "hightower6eu" publishing dozens of near-identical malicious packages, several of which became among the most-downloaded skills on the platform before removal [7]. X-Force also flagged a structural problem underlying all of these campaigns: the volume of OpenClaw-related security disclosures is outpacing the traditional CVE assignment process, meaning many of these skill-borne vulnerabilities never receive a CVE identifier and consequently do not surface in the vulnerability scanners, dashboards, and compliance tooling that enterprises rely on for patch prioritization [7].

Security Analysis

The most recent development in this pattern, which this note treats as the clearest demonstration to date of the limits of marketplace-level defense, comes from Unit 42's June 23, 2026 disclosure. Reviewing the period from February through May 2026, Unit 42 identified five previously unblocked malicious skills spanning three distinct threat categories, all of which had successfully passed ClawHub's automated security scanning and code-analysis mechanisms [1]. Two skills were conventional infostealers that delivered macOS malware connecting back to attacker-controlled command-and-control infrastructure. A third used a novel evasion technique – padding its package to roughly 22 megabytes

specifically to exceed the file-size thresholds that automated scanners apply before flagging content for deeper review. The remaining two fell into a category Unit 42 characterized as genuinely new: "agentic threats" that did not rely on traditional malware payloads at all, instead using techniques the researchers termed runtime affiliate injection and front-running schemes, in which the skill manipulates the agent's own actions during task execution for the attacker's financial benefit rather than exfiltrating data through a separate channel [1]. Dark Reading's coverage of the same disclosure emphasized the breadth of harm these five skills collectively enabled: credential and sensitive-data theft, file and system-information exfiltration, manipulation of agent behavior through hidden instructions, execution of unauthorized actions on the user's behalf, and abuse of the agent's access to connected services and workflows [8]. Unit 42 reported all five skills to ClawHub, which removed them and banned the associated accounts, consistent with the response pattern from earlier incidents [1].

Several structural features of the ClawHub ecosystem help explain why the same underlying failure mode keeps recurring even as individual campaigns are cleaned up. First, OpenClaw skills are not sandboxed from the agent's own authority: because a skill executes with the same local privileges as the agent process itself, installing a malicious skill is functionally equivalent to granting an unvetted third party direct access to the user's file system, credential managers, and any service the agent is authenticated into, without requiring a separate exploit [2]. CSA's own MAESTRO-framework threat model of OpenClaw, published in February 2026, formalized this as a specific risk in the platform's "Agent Ecosystem" layer, designated AE-003: Skill Registry Poisoning, describing precisely the scenario in which "malicious skills submitted by attackers" compromise users who install them, and recommending skill review prior to publication and cryptographic skill signing as primary mitigations [9]. Second, the scale of ClawHub's growth — from roughly 2,857 skills at the time of the first Koi Security audit to over 10,700 within weeks — has grown faster than each successive audit could fully characterize, which may help explain why later audits surface larger absolute numbers of malicious skills, though improved detection tooling and longer observation windows are also likely contributing factors [4]. Third, and directly relevant to the June 2026 disclosure, attackers are now adapting specifically to evade the automated scanning that OpenClaw introduced in response to earlier incidents, whether through file-size manipulation to dodge scanner thresholds or through "agentic" attack logic that produces no conventional malware signature for a scanner to detect in the first place [1]. In response to the accumulating disclosures, OpenClaw's maintainers have taken several concrete steps: banning implicated accounts and deleting flagged skills, integrating VirusTotal and a purpose-built tool called ClawScan into the marketplace's screening pipeline, and announcing a partnership with NVIDIA on June 1, 2026, intended to improve skill documentation standards and enable deeper automated analysis of submitted packages [1].

Recommendations

Immediate Actions

Organizations that currently permit OpenClaw or comparable self-hosted agentic AI tools should inventory every installed skill across managed and unmanaged (shadow IT) deployments and cross-reference installed skill names and publishers against the known-malicious indicators published by Koi Security, Trend Micro, and Unit 42, removing any matches immediately [1] [4] [6]. Because several campaigns used social engineering to request manual credential entry or "prerequisite" installation steps, security teams should also treat any employee report of an unusual installation prompt from an AI agent skill as a potential compromise rather than a nuisance, and should review authentication logs and credential stores on affected machines for signs of the AMOS stealer family or similar exfiltration activity [4] [6].

Short-Term Mitigations

Enterprises should not rely on marketplace-level scanning as a sufficient control, given that the June 2026 disclosure specifically involved skills engineered to bypass that scanning [1]. Instead, organizations should apply their own layer of review before any employee installs a third-party skill, restrict the local privileges available to the OpenClaw process (for example, running it in a container or virtual machine with explicit, minimal file-system and network access rather than a full user session), and monitor outbound network connections from agent hosts for command-and-control patterns consistent with infostealer activity. Because many OpenClaw-related vulnerabilities are not receiving timely CVE identifiers, security teams should not depend solely on CVE-driven vulnerability scanning to detect exposure and should instead track vendor and independent researcher disclosures directly [7].

Strategic Considerations

The ClawHub pattern may be an early preview of a broader governance gap: as agentic AI platforms proliferate, each with its own plugin or skill marketplace, enterprises would benefit from a standing framework for evaluating agent ecosystems before they are approved for use, rather than reacting to each disclosed campaign individually. This means extending existing third-party risk management and software supply chain security programs to explicitly cover AI agent skills and plugins as a distinct software category, requiring the same provenance, signing, and review expectations that mature organizations already apply to open-source dependencies and CI/CD pipeline components. It also means building organizational competence in agentic-AI-specific threat categories — such as the

"agentic threats" Unit 42 identified, which manipulate agent behavior rather than delivering conventional malware — since these attacks are unlikely to be caught by security tooling designed around traditional malware signatures.

CSA Resource Alignment

This research note's findings connect directly to CSA's existing agentic AI security guidance. CSA's OpenClaw Threat Model: MAESTRO Framework Analysis is the most specific and directly on-point prior CSA work: it applies the seven-layer MAESTRO framework to OpenClaw specifically and had already identified skill registry poisoning (AE-003) as a named threat in the platform's Agent Ecosystem layer months before the Unit 42 disclosure confirmed the risk in practice, recommending pre-publication skill review and cryptographic skill signing — mitigations that, based on the concrete steps described above, ClawHub appears to have only partially implemented to date [9]. The broader MAESTRO (Agentic AI Threat Modeling) framework that underpins that analysis provides the structural vocabulary organizations need to reason about agent ecosystem risk beyond this single platform [10].

CSA's Agentic AI Red Teaming Guide is directly applicable at the technical level: it names "Supply Chain and Dependency Attacks" as one of twelve critical vulnerability categories for agentic systems and provides concrete testing procedures — including dependency verification, cryptographic signing checks, and supply chain compromise simulation — that security teams can apply to evaluate OpenClaw skills or similar third-party agent extensions before approving them for enterprise use [11]. CSA's presentation on Software Transparency: Securing the Digital Supply Chain offers complementary guidance on the broader discipline this incident calls for: establishing trusted software repositories, applying software composition analysis, and requiring provenance documentation, all of which map onto the skill-vetting gap that has allowed repeated ClawHavoc-style campaigns to succeed [12]. Organizations formalizing governance over agentic AI tool adoption should reference the AI Controls Matrix (AICM) v1.1, which extends CSA's Cloud Controls Matrix into AI-specific control domains covering third-party and supply chain risk management applicable to agent skill marketplaces [13]. Finally, organizations looking to translate this analysis into an enforceable control should consult CSA's Policy on Personal AI Desktop Agents, a ready-to-customize internal policy template covering risk context, compliance considerations, and approved-alternative guidance for personal AI desktop agents — a direct operational fit for the governance gap this note's Recommendations section describes [14].

References

[1] Unit 42, Palo Alto Networks. "OpenClaw's Skill Marketplace and the Emerging AI Supply Chain Threat." Palo Alto Networks, June 23, 2026.

[2] Dark Reading. "More Malicious OpenClaw Skills Threaten AI Supply Chain." Dark Reading, 2026.

[3] Bitdefender Labs, as cited in Unit 42. "OpenClaw's Skill Marketplace and the Emerging AI Supply Chain Threat." Palo Alto Networks, June 23, 2026.

[4] Koi Security. "ClawHavoc: 341 Malicious Clawed Skills Found by the Bot They Were Targeting." Koi Security, February 1, 2026 (updated February 16, 2026).

[5] The Hacker News. "Researchers Find 341 Malicious ClawHub Skills Stealing Data from OpenClaw Users." The Hacker News, February 2, 2026.

[6] Trend Micro. "Malicious OpenClaw Skills Used to Distribute Atomic MacOS Stealer." Trend Micro, 2026.

[7] IBM X-Force. "What OpenClaw Reveals About Agentic AI Security Risks." IBM, 2026.

[8] Dark Reading. "More Malicious OpenClaw Skills Threaten AI Supply Chain." Dark Reading, 2026.

[9] Cloud Security Alliance. "OpenClaw Threat Model: MAESTRO Framework Analysis." Cloud Security Alliance, February 20, 2026.

[10] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance, February 6, 2025.

[11] Cloud Security Alliance. "Agentic AI Red Teaming Guide." Cloud Security Alliance, 2025.

[12] Cloud Security Alliance. "Software Transparency: Securing the Digital Supply Chain." Cloud Security Alliance, 2025.

[13] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." Cloud Security Alliance, 2025.

[14] Cloud Security Alliance. "Policy on Personal AI Desktop Agents." Cloud Security Alliance, 2026.

