ON2IT AND AMS-IX PRESENT: THREAT TALKS

Threat Talks
The Automous AI Risk Era

threat-talks.com

When intelligent assistants become operational attack surfaces

AI is no longer just generating text. It is executing tasks. Browsing websites. Running terminal commands. Sending emails. Managing credentials. Acting independently for all of it.

That shift changes the risk model entirely. The danger is no longer "wrong answers". It is excessive privilege combined with autonomy. When an AI agent has persistent memory, system-level access, and the ability to act without human validation, a single vulnerability can compromise an entire digital environment.

OpenClaw illustrates what happens when convenience outpaces guardrails. It is not the individual technologies that create risk. It is their consolidation into one continuously running, highly privileged AI operator.

In this Threat Talks infographic we discuss the following threats:

- CVE-2026-25253
- OpenClaw

87%
of global organizations experienced an AI-enabled cyberattack in 2025.

Source: AllAboutAI

42,000
More than 42,000 OpenClaw instances are exposed on the public internet.

Source: CSO

22%
of organizations have employees actively using autonomous AI agents without IT approval.

Source: CSO

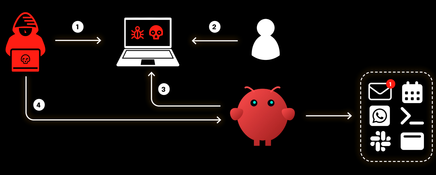
85%
of cybersecurity professionals attribute the increase in cyberattacks to generative AI.

Source: CFO

CVE-2026-25253 (OpenClaw vulnerability)

One-Click Remote Code Execution Through WebSocket Origin Bypass

CVE-2026-25253 is a critical Remote Code Execution vulnerability in OpenClaw caused by missing WebSocket origin validation. A victim only needed to click a malicious link for their browser to silently establish a connection to the locally running OpenClaw instance. Because the server performed no authentication or origin checks, the attacker gained operator-level control, allowing them to disable safeguards and execute arbitrary commands on the host machine. Given OpenClaw's access to terminal, browser sessions, messaging apps, and stored credentials, exploitation effectively meant full system compromise.



1. Malicious server
The attacker sets up a malicious server and crafts a link (e.g., via phishing email, social media, or a compromised website) designed to lure the user into clicking it.

2. WebSocket connection
When the user clicks the link, their browser immediately opens a page that silently initiates a WebSocket connection back to the locally running OpenClaw server. No additional user interaction is required: this is a true one-click exploit.

3. Operator-level access
Because the WebSocket server lacks origin validation and authentication, the attacker's page hijacks the connection and gains operator-level access. From here, the attacker can disable user confirmation prompts, turn off sandboxing, and switch OpenClaw into a mode that executes commands directly on the host machine, even if OpenClaw was originally running inside a container.

4. Full remote execution
The attacker now has full remote code execution on the user's machine. Given OpenClaw's broad system access (mail, terminal, browser, files), this effectively means complete control over the user's computer and all connected accounts.

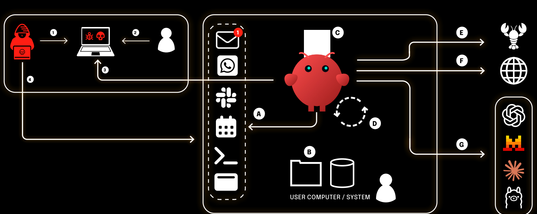
Mitigation

Upgrade to OpenClaw version 2026.1.29 or later, which adds WebSocket origin validation. If you ran a vulnerable version, rotate all credentials and API keys that OpenClaw had access to, and review system logs for signs of unauthorized command execution.

OpenClaw

An Autonomous AI Agent With Full-System Privilege

OpenClaw is an open-source AI agent that combines large language models, browser automation, terminal access, messaging APIs, and persistent memory into a continuously running assistant on a user's machine. Unlike traditional chatbots, it operates autonomously, retains long-term context, and has direct access to email, messaging platforms, files, and system commands. Compounding the risk, sensitive data such as API keys and passwords were stored in plaintext, and third-party skills could be installed from a public marketplace. The result is an AI system whose compromise extends far beyond a single application, reaching the user's entire digital environment.



A. Full access
OpenClaw has full access to the user's computer and its associated tools, including: email, messaging apps (WhatsApp, iMessage, Slack), calendar, terminal, and the web browser. This broad access is what makes it powerful and what makes a compromise devastating.

B. Data storage
OpenClaw stores all data in plain text files, including passwords, API keys, OAuth tokens, and credit card information. Making any file-level access a full credential compromise.

C. Persistent memory
One of OpenClaw's key features is its persistent memory. Every conversation and session is retained, which may include sensitive information such as internal documents, credentials shared in chat, or personal data.

D. Continuous loop
OpenClaw runs autonomously in a continuous loop. Without being prompted, it proactively checks for tasks it can perform on the user's behalf. Also meaning malicious instructions injected into its context can be acted upon without human oversight.

E. Extended capabilities
ClawHub.ai is a community marketplace containing skills that extend OpenClaw's capabilities. OpenClaw can discover, download, and activate these skills autonomously. Security researchers identified 341 malicious skills on ClawHub as part of the "ClawHavoc" campaign, including credential stealers and reverse shells disguised as legitimate tools.

F. Powerful automation
OpenClaw can control the user's web browser, enabling it to navigate websites, fill in forms, and perform transactions. While this enables powerful automation (like ordering groceries), it also means an attacker who compromises OpenClaw can perform actions on any website the user is authenticated to.

G. LLM-agnostic
OpenClaw is LLM-agnostic and works with various large language models, allowing users to choose their preferred provider. However, this also means API keys for multiple AI services may be stored in OpenClaw's plain text configuration.

threat-talks.com

ON2IT ZERO TRUST INNOVATORS

amsix

Contact us @
team@threat-talks.com