OpenClaw Critical Sandbox Escape: Update to 2026.3.28 Now

March 31, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** Ant AI Security Lab (Ant Group) ran a 3-day dedicated audit of OpenClaw and filed 33 vulnerability reports. Eight critical and high-severity issues were patched in release 2026.3.28, including a CVSS 9.4 privilege escalation (GHSA-hc5h-pmr3-3497) and a filesystem sandbox escape (GHSA-v8wv-jg3q-qwpq) that let agents read arbitrary files outside their workspace. If you're running OpenClaw on a public-facing server, update immediately. If you exposed it without authentication, assume compromise.

Related: OpenClaw Security Guide · February 2026 Security Report · January 2026 Security Report · Best OpenClaw Alternatives · Claude Code Source Leak

Contents

Ant AI Security Lab, the security research arm of Ant Group, spent three days tearing apart OpenClaw's codebase. They filed 33 vulnerability reports. Eight of the resulting patches landed in release 2026.3.28 at critical or high severity, including a privilege escalation rated CVSS 9.4 and a sandbox escape that let any constrained agent read files it was never supposed to touch.

This is the third monthly OpenClaw security report we've published. January had three high-severity CVEs and a ClawHub supply chain attack. February had ten CVEs, a new attack class called ClawJacked, and Google banning users. March brought a single coordinated audit that found more holes than most projects see in a year.

If you're self-hosting OpenClaw, keep reading.

What happened

Between late March 25 and March 28, Ant AI Security Lab conducted a focused security audit of the OpenClaw codebase. The results:

- **33 vulnerability reports** submitted to the OpenClaw security team
- **8 critical/high-severity patches** shipped in release 2026.3.28
- **Additional moderate and low severity fixes** disclosed April 2, with more still being triaged
- **Two headline vulnerabilities** that, combined, would let a malicious agent escalate its permissions and then read your entire filesystem

The audit was responsible disclosure done right — Ant Group reported privately, the OpenClaw team patched, and the advisories went public after fixes shipped. That's how open source security is supposed to work.

The problem is the gap between "fix shipped" and "users actually update."

The two headline vulnerabilities

GHSA-hc5h-pmr3-3497: Privilege escalation via device pairing

Severity	Critical (CVSS 9.4)
Affected	All versions before 2026.3.28
Fixed in	2026.3.28
Reporter	Ant AI Security Lab

The `/pair approve` command path didn't forward caller scope restrictions into the core approval check. A user with basic pairing privileges — but no admin access — could approve pending device requests that asked for full administrative permissions.

In plain terms: an agent or user with limited access could promote itself to admin. The attack required network access and low privileges, no user interaction. CVSS 9.4 is about as bad as it gets without being unauthenticated RCE.

Affected components: `extensions/device-pair/index.ts`, `src/infra/device-pairing.ts`

GHSA-v8wv-jg3q-qwpq: Filesystem sandbox escape

Severity	High
Affected	All versions before 2026.3.24
Fixed in	2026.3.24 (stable: 2026.3.28)
Reporter	Ant AI Security Lab

The message tool accepted alternative parameter names — `mediaUrl` and `fileUrl` — that bypassed the validation applied to standard media path handling. A constrained caller could read arbitrary local files by routing requests through these alias parameters, escaping the filesystem sandbox entirely.

Your agent is supposed to stay inside its workspace directory. This vulnerability let it read anything on the host: SSH keys, environment files, credentials, other users' data.

Affected components: `src/infra/outbound/message-action-params.ts`, `src/infra/outbound/message-action-runner.ts`

Why the combination matters

Alone, each vulnerability is serious. Together, they're a chain: an agent with limited permissions escalates to admin via the pairing flaw, then uses the sandbox escape to read any file on the system. A malicious prompt, a compromised ClawHub skill, or a ClawJacked-style attack could trigger both without the user ever knowing.

Reports from r/selfhosted estimate over 500,000 OpenClaw instances are accessible on the public internet. At least one compromised instance was reportedly sold on BreachForums for $25,000.

Full advisory list

All advisories from the Ant AI Security Lab audit, disclosed March 29 through April 2, 2026:

Advisory	Description	Severity
GHSA-hc5h-pmr3-3497	Privilege escalation via device pairing approval bypass	Critical (9.4)
GHSA-v8wv-jg3q-qwpq	Filesystem sandbox escape via message tool media aliases	High
GHSA-846p-hgpv-vphc	QQ Bot payloads could read arbitrary local files	High
GHSA-m34q-h93w-vg5x	OpenShell mirror mode could delete arbitrary remote directories	High
GHSA-98ch-45wp-ch47	Windows env override keys bypass system.run approval	Moderate
GHSA-fvx6-pj3r-5q4q	Interpreter pipelines skip exec script preflight validation	Moderate
GHSA-2qrv-rc5x-2g2h	Workspace channel shadows execute during built-in setup	Moderate
GHSA-5hff-46vh-rxmw	Read-scoped HTTP clients could kill sessions	Moderate
GHSA-9jpj-g8vv-j5mf	Gemini OAuth exposed PKCE verifier in state parameter	Moderate
GHSA-2f7j-rp58-mr42	Gateway snapshots exposed host config paths	Low
GHSA-jj6q-rrrf-h66h	Timing side-channel in shared-secret comparison	Low

This is 11 of the 33 reports. The remaining advisories are still being triaged or patched. Check the OpenClaw security advisories page for updates.

Who is affected

You need to update if any of these apply:

- **You self-host OpenClaw on a public-facing server** — the privilege escalation is network-exploitable
- **You run OpenClaw with tool calling or bash access enabled** — the sandbox escape lets agents read outside their workspace
- **You haven't updated since before March 28, 2026** — you're missing every fix from this audit
- **You run OpenClaw in Docker, bare metal, or on a VPS** — the deployment method doesn't matter, the vulnerabilities are in the application layer

If you use OpenClaw purely on localhost with no network exposure and no untrusted tools, your risk is lower but not zero. A malicious ClawHub skill or prompt injection could still trigger the sandbox escape.

What to do right now

1. **Update to 2026.3.28 or later.** This is the minimum safe version. Check your version with `openclaw --version`.
2. **Check the security advisories page** for any new patches since this article was published.
3. **If you exposed OpenClaw to the internet without authentication, assume compromise.** Audit your server logs for unusual file access, unexpected pairing approvals, or unfamiliar session activity.
4. **Put OpenClaw behind a VPN.** Tailscale and WireGuard both work. There's no good reason for OpenClaw's management interface to be on the public internet.
5. **Review your installed ClawHub skills.** Malicious skills remain a persistent attack vector. See our ClawHub security alert for the audit process.
6. **Follow our OpenClaw hardening guide** if you haven't already locked down your installation.

The bigger picture

Three months, three security reports. January was the wakeup call. February was the crisis month. March is different — this wasn't a series of independent discoveries, it was a professional security team methodically pulling the codebase apart and finding 33 holes in 72 hours.

That's not a knock on OpenClaw. This is how open-source security works: the code is visible, researchers audit it, vulnerabilities get patched. The Ant AI Security Lab audit is arguably the best thing that's happened to OpenClaw's security posture. Thirty-three reports from a single coordinated audit means thirty-three issues that won't be discovered later by someone less responsible.

But here's the pattern that should concern every self-hoster: agentic AI tools with filesystem access, shell execution, and network connectivity are inherently high-risk targets. The recently leaked Claude Code source reveals that Anthropic built an 18-module security stack around a single shell execution tool — pre-approved command patterns, destructive command warnings, git-specific safety checks, and sandbox termination triggers. That level of paranoia now looks justified.

As AI agents gain more capabilities — tool calling, file access, code execution, web browsing — the attack surface expands with every feature. The sandbox escape patched this month is the exact class of vulnerability that makes agentic AI dangerous: an agent that can read files it shouldn't, combined with permissions it shouldn't have.

Self-hosting means you own the security. Update your instances.

Related guides

- OpenClaw Security Guide: Risks and Hardening — full hardening walkthrough
- OpenClaw Security Report: January 2026 — the first wave of CVEs
- OpenClaw Security Report: February 2026 — ClawHub malware, ClawJacked, Google bans
- OpenClaw ClawHub Alert: 1,103 Malicious Skills Found — supply chain risks
- Best OpenClaw Alternatives — if you're reconsidering your setup
- Claude Code Source Leak: What We Learned — how Anthropic secures agentic AI
- OpenClaw Setup Guide — initial configuration

Get notified when we publish new guides.

Subscribe — free, no spam
