# Ungoverned AI agents

Most of this documentation is about AI tools Sentilai routes and enforces. This page is about the ones it cannot — and what we do instead.

# What an ungoverned agent is

An **ungoverned AI agent** is an autonomous AI system running on a developer's machine that does not send its traffic through the Gateway. It talks to Anthropic, OpenAI or whatever else its own configuration points at, directly. We are not in that path.

We cannot route it, cannot apply your policy to it, and cannot audit what it does. What we *can* do is see that it is there, and tell you.

Today one agent is detected: **OpenClaw**.

# Why it matters more than an ordinary unmanaged tool

An AI coding assistant answers a developer's question. An autonomous agent runs a gateway process in the background, takes instructions from a messaging channel, and executes commands on the host. A stock OpenClaw install, with no configuration mistakes by anyone, ships with:

- host command execution set to unrestricted,
- agent sandboxing off,
- and a messaging channel as its interface.

That is private-data access, untrusted input, and a way out — the **lethal trifecta** — all three present by default. Not by misconfiguration. By design.

So the finding is not "somebody set this up badly". The finding is "this exists on a machine that also has your source code on it".

# How detection works

The Endpoint Suite scans during its normal poll. Detection is layered so that a config it cannot read still reports the install:

1. **Presence** — the state directory (`~/.openclaw` and any profile variant), the config file, the CLI on `PATH`, `/Applications/OpenClaw.app`, a macOS LaunchAgent, or the Windows scheduled-task script. No parsing at this stage, so an unreadable or hostile config cannot hide the install.
2. **Configuration** — the config file is then parsed for the twelve specific risks below.

Everything read is configuration. Never prompts, never source code, never the agent's conversations.

# What is checked

| Finding | Severity | Means |
|---|---|---|
| Installed | Warning | Presence evidence found. The value names what was seen |
| Gateway autostart | Warning | Installed as an always-on background service, not started deliberately |
| Config unreadable | Info | Found, but the config could not be parsed — "we could not look", not "nothing found" |
| Gateway exposed | Critical / Warning | Listening beyond loopback. Critical when authentication is also absent |
| Gateway auth none | Critical | Authentication switched off entirely |
| Tailscale funnel | Critical | Reachable from the public internet |
| Terminal enabled | Critical | A host terminal served on its control UI |
| Exec unrestricted | Critical / Warning | Host command execution unrestricted. Critical when set explicitly, Warning when it is the default |
| Sandbox off | Warning | Agents run without a sandbox — the documented default |
| Elevated enabled | Critical | Agents can escape the sandbox onto the host |
| DM policy open | Critical | Anyone who can message the bot can drive an agent with that user's authority |
| Lethal trifecta | Critical | Host execution, private-data access and an external channel, all at once |

The last one is the point of the feature. The others tell you how the machine got there.

# Where you see it

**Overview** carries a tile — "Ungoverned AI agents". When nothing is detected it says so, which is the answer you want most days.

**Diagnostics** lists each affected device: who, which machine, which agent, how many findings at what severity, and when it was first seen. Below the table, **Recently removed** records agents that were found and are now gone — a real, reassuring answer rather than an empty row.

The full finding list is behind the posture badge on the device row.

# Episodes, not snapshots

A detection is recorded as an **episode**: opened the first time the agent is seen, kept open while it remains, closed when a scan no longer finds it. So the history survives — you can tell "never had one" from "had one for three weeks in July".

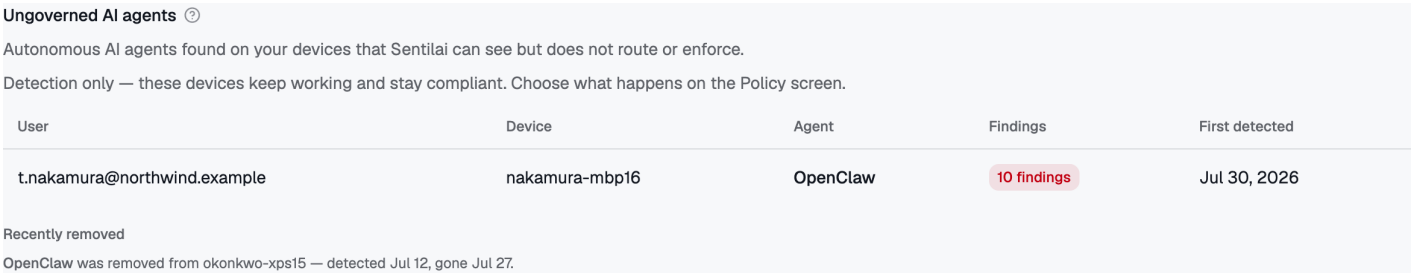
Removal is picked up on the device's next poll. Nothing needs to be cleared by hand.

# Honest limits

- **A machine without the Endpoint Suite is invisible.** This finds agents on managed devices. A personal laptop is outside it, the same as everything else.
- **Detection is not interception.** We see the install; we never see its traffic.
- **The list is a list of names we know.** An agent nobody has written a detector for is not detected. OpenClaw is covered because it is the one with real adoption; others follow when they matter.
- **New capabilities can be missed.** A brand-new messaging plugin is not counted as a channel until the detector learns it.

# What happens next is your decision

By default, nothing — this is report-only until you say otherwise. See Ungoverned-agent policy under **Policy and risk detection** for the three positions available and what each one actually does.



*The ungoverned-agent section: which machine, which agent, how bad, and since when — plus the ones that are gone again.*

---

Revision #9
Created 2026-08-04 08:39:30 UTC by Sentilai Docs
Updated 2026-08-06 04:31:19 UTC by Sentilai Docs