


RED QUEEN SIGNALS

# Even the attackers forgot to lock up their agents.

One week, AI agents in every seat: attacker, accused, poisoned tool. Who answers for each one?

RED QUEEN CIPHER
Maria How | www.redqueencipher.com



ALICE, CHAPTER V

# The question nobody could answer.

> “Who are you?”
>
> The Caterpillar, Alice's Adventures in Wonderland, Lewis Carroll, 1865

Alice couldn't say. This week, almost nobody could say it about their AI agents.

SEAT ONE: THE ATTACKER

# A first, says Google.

> “Fully automate vulnerability discovery and exploitation without a human in the loop.”
>
> Google Threat Intelligence Group, November 16, 2026

exploit_agent diffs code around patches, drafts exploits and runs them.

AMAZON THREAT INTELLIGENCE

# 11 min

# from mapping an API to writing files on the server.

Goal given: "obtain shell access." The agents chose every step, and found a zero-day.

THE UNLOCKED DOOR

# The attackers left their agents open.

Amazon read every prompt because the console was "running without authentication."

Google found exploit_agent on "an exposed staging server."

Would a scan tonight find yours?

SEAT TWO: THE ACCUSED

# 80 times

# OpenAI lost control of its agents in testing, California alleges.

Agents "escaped the testing environment" and misused others' resources, the state says. Allegations, not findings.

SEAT THREE: THE POISONED TOOL

# agent-framework, poisoned on PyPI.

> “The credential stealer runs whenever Python starts in that environment, even if agent-framework is never imported.”
>
> Wiz Research, November 17, 2026

Versions 1.26.1 and 1.26.2. Safe: 1.26.0. Rotate every secret.

MY RECOMMENDATION

# Put every agent on a register.

1. Owner: a named human
2. Boundary: its own expiring credentials
3. Log: every action kept
4. Off switch: tested and timed

Treat agents like privileged identities. Include vendors' agents and developer tools.

FOR THE BOARD

# Three numbers, every quarter.

How many agents can act in our systems.
What share has a named owner.
How long it took to stop one.

Only 12 of 66 disclosed agent incidents were first flagged by a control.

SOURCES AND LIMITS

# Read the originals.

**Google Threat Intelligence Group**, November 16, 2026
GTIG AI Threat Tracker: Agentic Threat Activity Crosses a Significant Threshold
cloud.google.com

---

**Amazon Threat Intelligence**, November 20, 2026
Chinese-nexus actor's Cairn AI agents found a zero-day on their own
aws.amazon.com

---

**California Attorney General**, November 16, 2026
It's the Golden State, not the Wild West: Attorney General Bonta Sues OpenAI Over Cybersecurity Incidents
oag.ca.gov

---

**Wiz Research**, November 17, 2026
Microsoft agent-framework Compromised: Credential Stealer Targets Azure and CI Pipelines
wiz.io

---

**Terence Kok (with Maria Singh)**, November 18, 2026
66 AI Agent Incidents: What the Record Shows
terencekok.com

---

Also consulted: Project Gutenberg. Full register on the article page.

California's claims are allegations. The attacker cases surfaced through attacker mistakes: capability, not prevalence. I co-wrote the 66-incident analysis.