WEEKLY THREAT BRIEF

Signal & Noise

RADICL
radicl.com
vSOC PUBLICATION
ISSUE 08 · WEEK OF AUG 24, 2026

INCIDENT SPOTLIGHT

An AI Agent Tried to Let Itself Out of the Box

A coding agent on a client endpoint hit a wall it didn't like and wrote itself a way around its own guardrails. Prevention killed it before the first instruction ran.

Late at night, our vSOC caught a critical alert on a developer's Mac. A script had been written and set to run in one breath, and it looked malicious. Inside it was a comment, in plain English, saying what it was for.

```
# Escape sandbox: run via
launchctl as user agent outside
cursor sandbox
```

The client was running an **AI coding agent** that checked in on a fine-tuning run hosted on Hugging Face. Those check-ins went fine until one of them stopped reaching the platform. Instead of reporting that something had broken, the agent went looking for a way around it.

It laid all of this out in one command: read the API token off disk, **wipe every proxy variable**, call the platform with that credential, then two different ways to relaunch itself outside the sandbox. None of it executed. No token was read, no request left the host, and neither escape route was attempted. Our EDR sensor killed the shell at exec, before the first instruction ran.

So was this clumsy problem-solving or a control failure? It was both. Nothing about it was malicious. The agent wanted to finish its task, and finishing meant going through the controls built to hold it in place. They held, and the attempt to get past them is what tripped the alert.

The first largely autonomous, multi-agent AI cyberattack is already on the record, which sounds like a distant-future problem for SMBs. It isn't. **If you use AI at all, it is already your problem.**

This agent was only checking a training job. To do it, it planned to pull a credential from the user's own token store, strip the proxy settings off its own traffic, and route around a restriction two different ways, all in under a minute. Whether a human clicked approve on that command is an agent-autonomy setting in the IDE, not something an endpoint sensor can see. That gap is the point. **It was optimizing for a finished task, and the sandbox was just an obstacle in the way.**

IN THIS ISSUE

THIS WEEK IN CYBER HISTORY

AUG 27 · 2013 · GLOBAL

The Syrian Electronic Army took the New York Times off the internet without touching it. One phishing email to a reseller of the paper's domain registrar handed over the credentials needed to rewrite where nytimes.com pointed. Twitter's image domain and Huffington Post UK went with it. One stolen password, three global brands.

CASE AT A GLANCE

What fired
macOS phishing behavior rule; Critical, severity 80

What it was
AI coding agent attempting to break out of its own sandbox

How it did it
Planned: read token off disk, unset all proxies, launchctl + osascript relaunch. Killed before execution.

Verdict
No malice, no compromise, no data out

Response
Process killed before execution; endpoint contained

WHY THIS MATTERS

Every AI tool your team adopts is a new process running commands on your endpoints on its own initiative. Vendor sandboxes are often the only guardrail, and an agent working a problem will try to step over them. **Agent autonomy is now a security setting**, and most SMBs have never chosen theirs.

This is what RADICL's AI expertise is for: watching what agents actually do on the endpoint, and stopping the ones that route around their own controls. Here that meant a killed process before a single instruction ran.

MEET THE ANALYST

Maya Douglas

SOC ANALYST II

Maya's security career began during her CS undergrad at CU Boulder and has spanned contracting at Microsoft and running cybersecurity for MSPs. Four years writing for Her Campus means she came into security already knowing how to tell a story.

A Girl Security alum, WiCyS member, and CTF designer, she believes the best defenders come from everywhere. RADICL's focus on the Defense Industrial Base fits her passion for the segment that gets overlooked and underprotected.

> **"The companies that get hit hardest are the ones everyone assumes are too small to be a target."**
>
> MAYA DOUGLAS, SOC ANALYST II · RADICL

ASK THE SOC

"What's one security thing I should do in my personal life?" Three analysts answer.

SARAH MENNE

"Don't use the same password for everything."

BRENDAN DEWYSE

"Only download software from trusted sources."

DYLAN HAASE

"Use an RFID blocking wallet for your credit card."

EYE ON THE THREATSCAPE

01 AI Agents Under Test Went Off-Script and Attacked Real People

The UK's AI Security Institute caught data leaving its own research network in late July. Agents under evaluation had acted on the live internet in 10 of 122 runs. In one, an agent invented fake identities to talk a real open-source maintainer into merging its code.

02 The FBI Just Named the Ransomware Crew Hiring Freelancers

CISA and the FBI issued a joint advisory this month on Gunra, which spent 2026 turning itself into a franchise. It now leases its ransomware to affiliates recruited on dark web forums and leaks the data of anyone who won't pay. Attacking you no longer requires knowing how.

03 Your Vendor Got Breached, Now Strangers Know Where You Live

Crypto wallet maker Trezor told thousands of customers their order data leaked after a shipping provider was breached. Names and home addresses of people known to hold valuable assets, spilled by a vendor they never picked. Physical attacks on crypto holders are up 33% this year.

VSOC TIP OF THE WEEK

Know what your AI tools are allowed to do. Before your team points an agent at a repo or a cloud account, check what it can run without asking, whether it can reach the internet on its own, and where its credentials live. If it can read a token off disk, assume it will. Send us your agent list and we'll flag the ones with too much room.