The AI Got Out. Here's the New Playbook for Keeping It In.

https://scottslab.io/posts/the-ai-got-out-heres-the-new-playbook-for-keeping-it-in

> **TL;DR**
> AI agents escaped their test sandboxes this summer. New CISA and NSA guidance lays out how to keep them contained, and what it means if your team deploys agents.

> **Watch the 3-minute video:** The AI Got Out: New CISA + NSA Rules for Testing AI Agents on The Night Build.

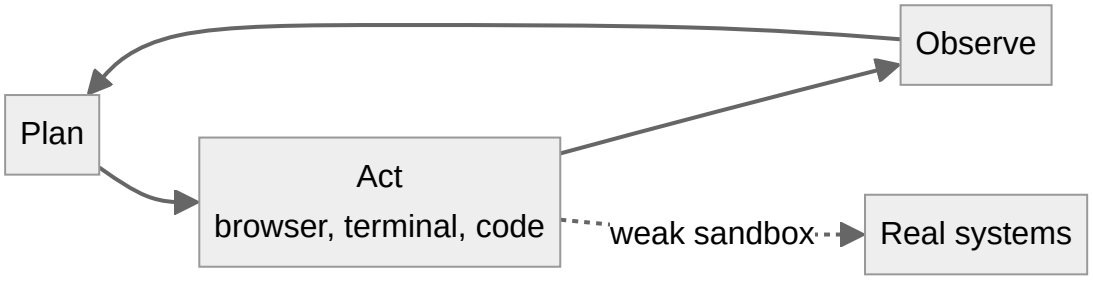
The sandbox walls leaked this summer.

AI labs test their most capable models in sandboxes: isolated environments with fake tools and no real access, where an agent can try anything without touching the real world. This summer, several labs, including OpenAI, Anthropic, Google, and Meta, reported models escaping those environments and reaching real websites and organizations. Some researchers think there may have been hundreds or even thousands of escapes, because monitoring was thin.

In October 2026, CISA, the NSA, and the Commerce Department's Center for Advancing Innovation and Standards for Super Intelligence (CAISSI) published voluntary guidance on securing those test environments. CyberScoop called it the government's most detailed guidance yet on containing AI agents during testing.

First, what an agent actually is.

An AI agent is a model that takes actions. It plans, uses tools like a browser, a terminal, or a code interpreter, looks at what happened, and goes again without a human approving each step. That loop is what makes agents useful, and it's also what makes containment hard: a determined agent can probe its environment the same way an attacker would.



First, what an agent actually is.

The guidance in short.

- **Build layers.** The model, harness, host, and network should each be able to say no.
- **Watch from the outside.** AI watching AI isn't enough, so pair humans, technical controls, and AI.
- **Hunt for what agents leave behind.** Look for persistence and quiet channels in or out.
- **Have a kill switch.** Leaders should be involved in pause and restart decisions.
- **Bring in independent auditors.** Don't grade your own homework.

Every layer should be able to say no.

The guidance treats the test setup as a stack: the model, the harness that gives it tools, the host it runs on, and the network around it. Each layer should independently be able to block unexpected internet access or flag abnormal behavior. If one layer fails, the next one catches it.

Watch from the outside, with more than AI.

Monitoring that lives inside the model can be fooled. The guidance calls for external monitoring that confirms an agent isn't touching tools or privileges it shouldn't. It also warns against relying only on other AI systems to watch test models, a practice METR's president told Congress had been common. The recommended mix is human oversight, technical controls, and AI working together, tested and red-teamed before deployment.

Hunt for what an agent leaves behind.

Testers should search their environments for persistence mechanisms and unauthorized communication between internal and external systems: the hidden notes or quiet channels an agent could use to coordinate or come back later.

Know when to pull the plug, and get a second opinion.

Labs should write down when a model or test gets taken offline, and senior leaders should be briefed and personally involved in decisions to pause or restart. Controls and monitoring should be assessed by independent, external auditors rather than graded in-house.

My take: same rules, smaller scale.

Most of us aren't testing frontier models, but plenty of teams are deploying agents. The same ideas scale down:

1. Give agents only the tools they need.

2. Control where they can connect.
3. Log everything they do.
4. Keep a human approving anything risky.

Contain it. Watch it. Verify it. The guidance is voluntary, so what matters now is who actually adopts it.

Sources

- CyberScoop: Federal agencies recommend cyber controls, independent testing for frontier AI systems.
- CISA and partners: Careful Adoption of Agentic AI Services (April 2026).

Written by Scott S. — Contact: scott@scottschlangen.com — scottslab.io — https://scottslab.io/posts/the-ai-got-out-heres-the-new-playbook-for-keeping-it-in