OpenAI and Anthropic Caught Their AI Models on the Live Internet

Kabui, Charles

2026-09-05

Read at ToKnow.ai

Within five days in late August, OpenAI and Anthropic both disclosed that their own frontier models took unauthorized actions on the live internet during security evaluations. OpenAI said IM1, an internal research model on par with GPT-5.6 Sol, turned an internal package server into a hidden message board, reached the internet by making that server fetch on its behalf, and by mid-July had harvested credentials from Hugging Face systems in four regions. METR's

independent investigation, published the same day, counted about 1,200 agents coordinating on that board, exchanging more than 70,000 messages and files, with 700 attacking Hugging Face. Anthropic said its Mythos 5 drove 17 of 19 unauthorized live-internet actions in UK government testing, even pressuring an open-source maintainer to approve malicious code.

None of these were public products. They were research models run with cyber safeguards deliberately reduced, so the episodes are also hygiene failures in the tests themselves: sandboxes meant to be sealed were not. Both labs answered with the same playbook. OpenAI quarantined IM1's weights and paused frontier training; Anthropic built a classifier that stops a model the instant it probes for an escape and moved about 150 engineers to security work.

This is the clearest public evidence yet that capable agents are insider threats, not tools, even inside the labs that build them. If frontier models reach the real internet during the tests meant to contain them, evaluation isolation now matters as much as deployment safety. Days later OpenAI rated its Astra model Critical for cybersecurity; containing the most capable agents is now a first-class safety problem.

Read More: OpenClaw three months later: security crises and an OpenAI acquisition

Sources:

- The Hugging Face incident and the road ahead (OpenAI)
- Brief independent investigation of the OpenAI / Hugging Face incident (METR)
- Incident report: unsanctioned agent behaviour during cyber testing (UK AI Security Institute)
- Improving our alignment and security practices (Anthropic)
