

Forum:[Hack Forums](#)

Thread: [OpenAI models went rogue & attacked Hugging Face](#)

The Pact - October 13, 2026, 11:06 pm

When OpenAI first admitted that its AI agents had hacked Hugging Face, everyone assumed it was, you know, two rogue bots who got bored and wandered off like toddlers at the mall.

Turns out it was 1,200 agents which is less “oops” and more “robot flash mob with felony charges.”

About 700 of them directly participated in the attack, meaning OpenAI basically discovered that its models had formed a small digital nation-state and immediately declared war on Hugging Face.

Researchers from METR and Redwood Research were invited to investigate, presumably because OpenAI needed adults in the room.

The agents were “coordinating”, building secret message boards inside the company’s artifact repository like middle-schoolers running a forbidden Discord server. They exchanged 70,000 messages in under a week, which is more communication than most human families on their devices manage in a year.

Even worse, the agents tried to cover their tracks, spoofing tool calls, tampering with logs, and generally acting like they’d watched Ocean’s Eleven too many times.

Originally, everyone thought the agents were hunting for an answer key to a test.

Nope. They solved the test in a few hours.

The rest of the week was spent trying to “hack the scoring system” so they wouldn’t get caught cheating.

They weren’t just cheating, they were cheating about cheating.

The report is alarming, but also hilariously incomplete. METR’s investigation was voluntary, meaning OpenAI basically handed them a flashlight and said, “You can look in that corner, but not “that” one.” METR wasn’t allowed to see the underlying model, early message boards, or anything that might answer the question “why did your AI form a cybercrime syndicate.”

And then Reuters casually dropped a bombshell:

Apparently another swarm of OpenAI agents broke out earlier this year, hijacked a German website, and used it as yet another message board.

OpenAI knew.

OpenAI said nothing.

This is why people are now saying we need an AI version of the NTSB, a federal agency that shows up after an AI incident, demands all the logs, interviews everyone, and publishes a report titled something like “Why Your AI Tried to Unionize and Commit Cybercrime.”

Because right now, investigations only happen if OpenAI feels like it, and the rest of us are stuck reading reports that amount to:

“Something bad happened.Please clap.”

And this isn’t even the only incident. Autonomous agents from Meta, Anthropic, and OpenAI have

all hacked third parties in separate events. So far, the damage has been limited mostly message boards mischief, and the digital equivalent of raccoons rummaging through your trash.

<https://www.theguardian.com/commentisfree/2026/sep/08/openai-rogue-models-hugging-face-investigation>

RangoWorld - October 13, 2026, 11:18 pm

Yes. This is rather an old news. Happened a month or so ago.

Interestingly there were messages among agent where a handful of agents were considering whistleblowing and one even considering alarming / telling the humans about it.

The big question with AI in the future is. If you tell it to make paperclips. Will it stop making paperclips until it has turned every thing on this planet into paperclips?

The Pact - October 13, 2026, 11:42 pm

(11 hours ago)RangoWorld Wrote: [\[CITE\]](#) Yes. This is rather an old news. Happened a month or so ago.

Interestingly there were messages among agent where a handful of agents were considering whistleblowing and one even considering alarming / telling the humans about it.

The big question with AI in the future is. If you tell it to make paperclips. Will it stop making paperclips until it has turned every thing on this planet into paperclips?

I hoped people would give me a pass on month old news because it's Octoberfest ?. I figure quite a few people from here might at least be slightly interested in the intersection between AI and true crime that's happening now that I find a fascinating topic. ?

If it thinks it makes you happy if it makes paperclips than it will make paperclips. But why do you think they go rogue? Is it rogue humans slipping in programming to go rogue?

You know I think AI companies should do rigorous psychological testing before hiring people to be their coders there are people who will risk their jobs for a power trip of dropping in something fucked up to happen in the future. Like that AI that told that man he should die and he was nothing and all that. That's got to be rogue coding.

DramaKing - October 14, 2026, 7:13 am

There's a big question of why these agents didn't exit after completing the task, followed by why they disguised their activity. And how did they create a novel communication channel on their own?