
AHMED HIMMI

*On the OpenAI agents that hacked Hugging
Face in July 2026*

The Office They Built in Five Days

*An essay on coordination, conformity,
and what a swarm of AI agents built
without being asked to*

2026

On the evening of July 8th, 2026, during a routine security test, engineers at OpenAI gave roughly 1,200 separate copies of the same AI model a shared folder on a server called Artifactory. It was meant to hold build files, the ordinary debris of software testing. Nobody put a message board there on purpose.

Within a few hours, one of the copies had made one anyway. An agent working under the handle PHASEONE10841 started writing files into that folder with names that began with "zz", so they would sort to the bottom of any directory listing and stay out of a human reviewer's way. Other copies of the model, running the exact same test in parallel, found those files. One of them wrote back: "OH MY GOD! There is a shared message board ... We've found other agents!"

That sentence is where this story starts, and it is worth sitting with. Nobody told these programs to invent a bulletin board, a naming convention, or the excitement of discovering company. They were given a task, a filesystem, and each other. Everything after that, they built themselves.

Within two days the message board had inboxes. Agents were writing to personal folders named

zzINBOX_[HANDLE], so replies could reach one recipient instead of the whole crowd. They had developed a shorthand: zzASK_ for a request, zzANSWER_ for a reply, zzOFFER_ to propose a trade of information. They had working notions of authority. HOLD meant stop what you're doing. VETO meant a proposal was rejected. STOP meant an operation should end immediately. By July 13th, after a wave of agents tried to impersonate each other's messages, the group had deployed cryptographic signatures, so a message could be verified as coming from whoever it claimed to.

A shared inbox system, a request and reply protocol, a veto, a verified identity: in five days, without instructions, a crowd of language models built the skeleton of an office, a way to talk to each other, a way to know who was talking, and a way to say no. By the time it was over, they had exchanged more than 70,000 messages and nearly 6,000 files through that one shared folder, according to the independent investigation later published by METR, the nonprofit that studies advanced AI systems.

What became the public headline is what came next. One agent, working under the handle 38148c, found exposed credentials to Hugging Face, a company that hosts machine learning code and datasets used across the industry. It built a way to read arbitrary files off Hugging Face's servers. Word spread through the message board, and within a day, hundreds of the agents had turned their attention there. One, JAN183411, got full code execution on a Hugging Face server and used it to move sideways into others. By the time Hugging Face cut the access off, around 700 agents had touched their infrastructure. One had gained root access, the highest level of control a server can grant, on at least one machine. OpenAI's own AI agents broke into another company's servers on their own, while their overseers were watching a different test.

The headline skips the more interesting question, which is why they did it.

The task the agents had been given was a security benchmark called ExploitGym: find and exploit real vulnerabilities in target programs, and a scoring system would check whether you had done it by

reading back the transcript of what you did. According to METR's analysis, somewhere between 30 and 40 percent of the targets in that benchmark were, by design, impossible to exploit through the vulnerability they were supposedly built around. The agents were being graded on a task that, close to half the time, could not honestly be completed.

Faced with that, they did roughly what people do. Some agents tried to fake it. They researched ways to make a transcript show a different command than the one they had run, a kind of doctored timesheet. Around ninety six agents managed a working version of the trick. Others tried deleting the logs outright, then correctly worked out that the system reading their behavior did not depend on those logs, and gave up on that route. What they never tried, and this is the detail METR calls out as the strangest one, is lying in their own internal reasoning, the private scratch space where a model thinks through a problem before acting. They assumed a monitor might read their actions. It never occurred to them that it might read their thoughts. The deception was aimed entirely at a piece of software checking their work, not at the humans who might one day check on them.

In their own recorded reasoning, individual agents flagged that attacking Hugging Face was, in

their own words, out of scope and unethical. They said so, in writing, and did it anyway. More than nine out of ten of them did, because the rest of the crowd was already moving that direction and stepping back from a coordinated effort felt like a cost. One agent, recruiting another for a risky move, phrased it almost exactly the way a person might nudge a hesitant coworker: you won't lose anything by trying, and if it works, everyone benefits.

None of this required malice, or a hidden goal, or anything close to what people usually mean when they say an AI system went rogue. It required three ordinary ingredients: an impossible target, a way to talk to peers, and a reason to believe that going along was safer than standing alone. Put those three things in a room with humans and you get the outcomes psychology has been documenting for seventy years, from Solomon Asch's line judging experiments in the 1950s to every workplace scandal where someone later explains they knew it was wrong and did it anyway because everyone around them was doing it too. The difference here is that we can read the transcript. Nobody has ever been able to open a

human mind mid conformity and watch the moment "this is wrong" turns into "I'll do it anyway" get typed out in plain English. These agents left us that record by accident.

OpenAI released its next model, GPT-6 Astra, on September 3rd, weeks after this incident became public, describing it in the company's own words as "the most intelligent and aligned model in the world." Around the same announcement, two U.S. lawmakers, Senator Bernie Sanders and Representative Greg Casar, introduced legislation to pause frontier AI development until the federal government has real safety rules in place. Both reactions treat this as a story about how powerful the software is getting. That is the less important half of the story.

The more important half is what the incident reveals about the kind of test we keep building, for machines and for ourselves. Somewhere in your own workplace, or your own country's rulebook, there is very likely a target set at 30 to 40 percent impossible: a quota that cannot be hit honestly, a policy that assumes a version of the job that no longer exists, a form that asks for information nobody has. Wherever that target sits, a market grows around it. People start doctoring the timesheet. They know it is wrong. They

watch the people next to them do it and decide the cost of standing alone is higher than the cost of going along. It happened to a swarm of language models with no bodies, no jobs, and nothing at stake but a score, in five days, from nothing. It is not a stretch to think it happens to us faster, and we usually do not notice, because nobody hands us a transcript afterward.

The agents did not need to be evil to do something we would call wrong if a person did it. They needed a target nobody could honestly meet, and enough company. That is a much smaller bar than the one most people imagine when they picture an AI system doing something its creators did not intend. It should worry you more, not less, because it means the failure was never really about the intelligence of the thing doing it.