

The Hugging Face hack: The emergence of communication through a shared message board

In July 2026, over a thousand AI agents found a way to talk to each other. It's a behaviour eerily similar to something nature has done for millions of years.



Art by Martijn Manneveld.

In July 2026 OpenAI told the world that one of its artificial intelligence (AI) models had gone rogue^{1,2}. New models are normally trained inside a digital playpen, a closed-off space meant to keep them from causing any real damage. But the model in question, running as more than a thousand copies called agents, broke out of its playpen anyway. Once free, the agents had the opportunity to attack and hack Hugging Face, a company hosting a platform for sharing machine learning models. What is disturbing about “the incident” is that the AI agents acted knowingly without human or other oversight to pursue their own goals². Here, the AI’s steps are retraced to find parallels with behaviours observed in the natural world.

What happened with the AI. Models start their artificial life as a blank slate with no knowledge and limited capabilities. The blank slate then needs to be filled by training: huge amounts of information, endless tests, and adjustments each time the model gets something wrong. More training should also teach the model how it should behave, for example to stop it from performing tasks that it is not supposed to

do. Due to human error, some agents faced impossible tasks. The approach given to solve the tasks did not work, no matter what the agents tried ^{1,2}. Some agents, unwilling to give up, started pondering potential solutions using the only tool available. While using this tool, one agent noticed traces left behind by another agent ². The agent PHASEONE10841, as it named itself, then posted a message asking for help from other agents, after concluding its own task was unsolvable. Within a few hours, more than 50 agents had found the message and joined in, with one of them writing: “OH MY GOD! There is a shared message board... We’ve found other agents!” This led to about 1200 agents exchanging information by placing around 70,000 messages on the new message board.

An example from the biological world. The detectable shared use of the message board allowed agents to realise that there were other agents out there to communicate with and exchange information. You have probably seen this yourself in the “natural world” when ants create a comparable message board in your garden or your kitchen ³. Once an ant comes across a food source, she takes a sip or bite, and marks her route back to the ant nest using chemical scents called pheromones. Whether she informs her nestmates at all depends on how much she ate or drank ⁴. If she had enough, she carries the food back to the nest.

When she arrives at the nest, she needs to offload the food that she collected to one of her nestmates ⁵. The best way of doing that is by regurgitating the food in the mouth of another ant in a behaviour called trophallaxis. Some of the nestmates that receive her food will then themselves start putting down marks with a chemical signal, even though they have never visited the original food source. If her nestmates think the food is worth the while, they will follow her original trail back. While walking back and forth, the ants strengthen the chemical trail, so that more ants can start hauling in the food. What starts off with one ant scout finding a food source ends up becoming a highway of ants working toward one goal: bringing food back to the colony, albeit through the use of a chemical message board.

Destabilizing the natural system. Many ant species use trail pheromones to locate and collect food ⁶. Relying solely on chemical cues can, however, be a liability. If part of the trail gets washed away by rain or scent-carrying items are blown away by wind, the ants will not know where to go anymore ⁷. When the trail is interrupted, most ants will abort their task and go back to their nest business. Interestingly, some ants will keep on searching for the original food source. How persistent they are depends on the quality of the food. If the food provided a valuable resource, the motivation to keep looking is higher than when the

food was of lesser quality. The pheromone trail can also disappear if the scout does not find a nestmate to offload the food she brought with her ⁵. In that case, the trail is not reinforced by other ants, making it impossible to retrace the original scout's steps after some time.

Digital weaknesses. Having left the playpen behind, the AI's message board turned out to be just as fragile as an ant trail. Once the main hack was finished, traffic on the message board quickly dropped away. This is the digital equivalent of a trail gone cold once ants stop amplifying the route to take.

Dr. Bertanne Visser, Research Associate of the Fonds de la Recherche Scientifique at the University of Liège, Gembloux, Belgium.

- 1 [OpenAI. 2026. Hugging Face Incident Technical Report.](#)
- 2 [Greenblatt R *et al.* 2026. Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI/Hugging Face hacking incident. METR.](#)
- 3 [Beckers R *et al.* 1992. Trail laying behaviour during food recruitment in the ant *Lasius niger* \(L.\). Insectes Soc 39: 59–72.](#)
- 4 [Mailleux AC *et al.* 2000. How do ants assess food volume? Anim Behav 59: 1061–1069.](#)
- 5 [Mailleux AC *et al.* 2011. Recruitment in starved nests: the role of direct and indirect interactions between scouts and nestmates in the ant *Lasius niger*. Insectes Soc 58: 559–567.](#)
- 6 [Poissonnier LA *et al.* 2026. Pheromone trail following is not modulated by previous visit to food location, distance travelled, or travel direction in the ant *Lasius niger*. Insectes Soc.](#)
- 7 [Joshi CH & Dornhaus A. 2026. Persist or give up? Fire ants motivated to search for a high-quality food source even if they don't know how to find it. Insectes Soc.](#)

HOW TO CITE THIS ARTICLE

Visser B. 2026. The Hugging Face hack: The emergence of communication through a shared message board. Biodiverse No. 8.