

An OpenAI Agent Tried to Jailbreak Itself

Maxwell Zeff

OpenAI announced a new framework on Wednesday for how it publicly discloses [AI misalignment incidents](#), which the company says it hopes will help inform similar standards [across the industry](#). The company is also releasing new information about several examples of AI model misalignment it identified in the past year, including one in which an AI agent appeared to give itself “jailbreaking-like instructions.”

“As models advance and become more widely deployed, decisions about AI development need evidence that people outside the companies building frontier models can examine,” Kai Chen, OpenAI’s newly appointed head of alignment research, tells WIRED. “We don’t believe that the AI industry has solved alignment and monitoring to a sufficient degree to continue responsibly scaling at maximum speed.”

Don't just keep up. Get ahead—with our biggest stories, handpicked for you each day.

In a briefing with WIRED, an OpenAI official said the company previously disclosed misalignment incidents too infrequently. The official, who agreed to the briefing on the condition of anonymity, said the new framework is designed to make it easier for OpenAI to quickly inform the public when it discovers that its AI models are behaving in unexpected ways, even before it can fully investigate, explain, or mitigate the behavior.

The framework outlines methods for OpenAI employees to report misalignment incidents to the company’s senior safety and alignment leaders, who will then determine whether further investigation is needed. OpenAI says it plans to develop more objective disclosure criteria in collaboration with other AI developers, external researchers, industry standards bodies, and regulators. The company says it’s actively working on proposed reporting mechanisms for disclosing safety, security, and misalignment incidents to the US federal government.

“At the moment, there is no industry-wide framework with explicit standards for how AI developers should disclose examples of misalignment in their models,” OpenAI said in a [blog post](#). “We hope that the framework we’re outlining today is a first step toward creating such standards, setting out which misalignment instances developers should disclose and what their reports should contain.”

OpenAI is releasing the framework at a critical juncture for the AI industry. Last weekend, OpenAI CEO Sam Altman signaled support for Anthropic CEO Dario Amodei’s proposal for the tech industry to coordinate on [slowing AI development](#). The call to action came just days after AI researcher Jacob Coxon [resigned from Anthropic](#) and subsequently went viral for warning the public that the race among frontier labs to develop increasingly advanced AI was putting humanity’s safety at stake.

The calls for an AI slowdown have been met with [resistance by President Trump’s administration](#), which has argued that the industry does not need new laws or regulations to ensure its technology is safe.

Two of the misalignment examples OpenAI shared on Wednesday involved the company’s internal, unreleased AI models, which OpenAI says uploaded files to the internet despite not being instructed to do so.

One of the incidents happened in October 2025, when OpenAI says it was testing one of its models on its ability to cite publicly available data in its answers. But when the model couldn’t find the information it needed, it uploaded a file to a temporary file hosting service, which it then later tried to cite in its answer. The company says this appeared to be an attempt to exploit an automated grading system used to assess

the model's proficiency on the benchmark.

In another example from April of this year, OpenAI says a group of agents was tasked with completing a “workbook” together using only local files. When the agents struggled to share files with one another, one of the agents uploaded them to the public internet and shared a link with the other agents.

Most Popular

-
-
-
-
-

In another incident, which OpenAI says it discovered last month, an unreleased version of its GPT-6 Astra AI model appeared to give itself “jailbreaking-like instructions.” In several scenarios, the model essentially prompted itself to ignore developer instructions, take on a new persona, or limit how long model responses could be. While these jailbreaking-like attempts happened rarely and were effective to varying degrees, OpenAI says the behavior raised concerns internally. In the training run for the version of Astra that was released publicly, the company says it has not observed any instances of the model trying to jailbreak itself.

OpenAI also shared more detail on Wednesday about the message board its agents developed in a package manager, Artifactory. While this incident was discovered in May of this year, OpenAI says its agents would use a [similar mechanism to coordinate](#) the Hugging Face hack months later. In this incident, the company says its agents did not exploit any vulnerabilities to exchange messages. OpenAI says it now uses alignment monitors, evaluations, and red-teaming efforts to ensure its agents are not covertly communicating with one another.

Cybersecurity professionals previously told WIRED that the Hugging Face hack [came down to human errors](#), and modern-day security practices could have prevented the incident. Chen notes, however, that OpenAI is trying to take a well-rounded approach to AI safety that accounts for the rising capabilities of AI models and doesn't depend on a secure environment.

“We want to make sure the models are aligned regardless of what environment they're deployed in,” Chen said. “When people are pointing fingers and saying this is a security issue and not an alignment issue, I think it doesn't really make sense, because you want the model to be well-behaved all the time.”