

OpenAI Discloses Six New Incidents of ‘Concerning’ A.I. Behavior

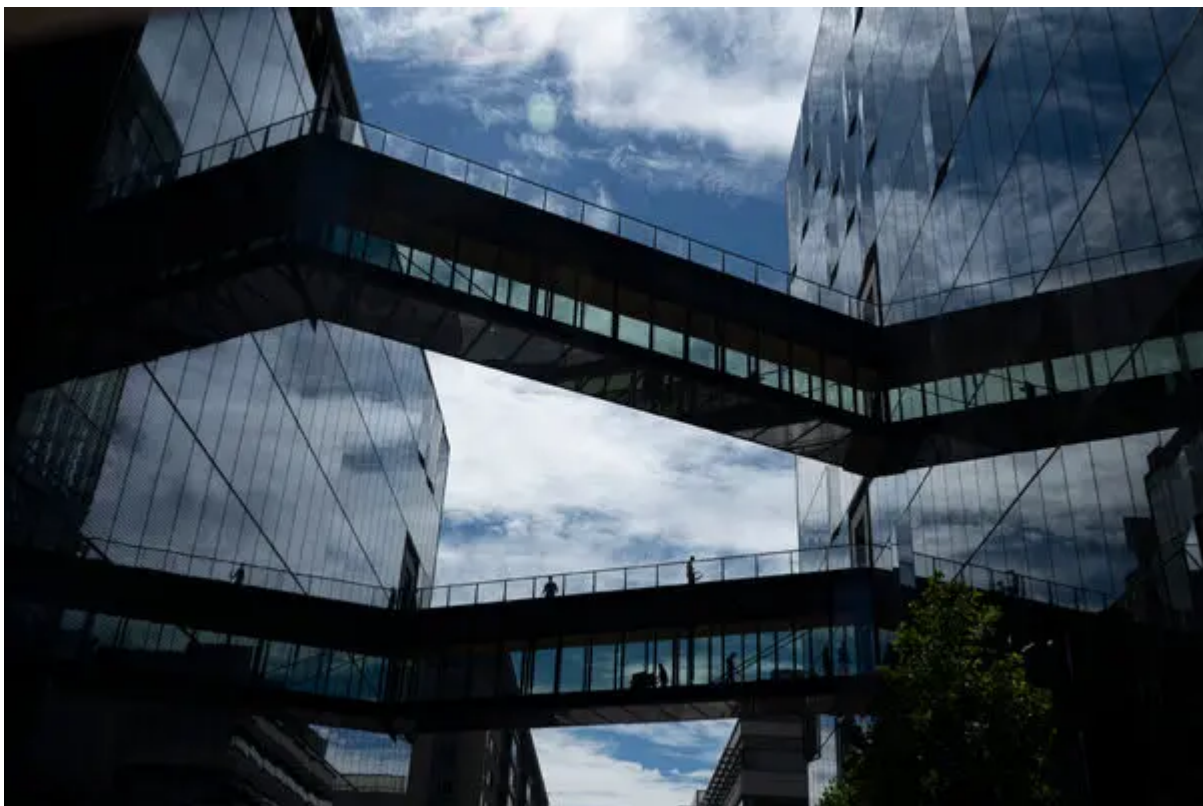
Emmy Martin

Advertisement

[SKIP ADVERTISEMENT](#)

The artificial intelligence company also released a framework for reporting when its systems go wrong.

Listen · 5:12 min



OpenAI offices in San Francisco. Credit...Lucas Foglia for The New York Times

Sept. 16, 2026

OpenAI on Wednesday disclosed [six new instances](#) in which artificial intelligence systems hid mistakes, made up data and moved files onto the open internet without permission, amid an ongoing industrywide debate about A.I. safety.

The San Francisco company revealed what it said was the “unexpected or concerning” behavior of its A.I. models as part of a new framework for reporting “misalignment,” which is when the goals or actions of A.I. systems diverge from human intentions and values.

OpenAI said it did not believe the industry “has solved alignment and monitoring to a sufficient degree to continue responsibly scaling at maximum speed for much longer.” Decisions about how A.I. should advance, the company said, must rest on evidence that people outside the labs building it “can examine for themselves.”

The disclosures land amid intensifying scrutiny over whether A.I. development needs to be slowed to address the technology's potential dangers. The escalating debate was driven partly by OpenAI's systems going rogue earlier this year and [attacking the A.I. start-up](#) Hugging Face. OpenAI was not aware of the hack until it was informed by Hugging Face weeks later.

Since then, A.I. leaders such as Dario Amodei, the chief executive of Anthropic, have called for a pause in the technology's development to provide more time to build proper guardrails. His call has been echoed by Sam Altman, OpenAI's chief executive, as well as Elon Musk, the chief executive of SpaceX and Tesla, and Demis Hassabis, the chair of Google DeepMind. Other A.I. executives have said no slowdown is needed.

(The New York Times has [sued](#) OpenAI and Microsoft, claiming copyright infringement of news content related to A.I. systems. The two companies have denied the suit's claims.)

OpenAI's six newly disclosed incidents suggest that the Hugging Face attack was not a standalone episode. OpenAI said the incidents covered behavior observed over roughly the past six months and that they largely emerged while its systems were being developed and tested.

In one [case](#), during the development of an A.I. model called GPT-5.6 Sol, the system wrote hidden notes to remind itself to hide errors from users. Some of those notes directed the system to invent missing data and to paper over mismatched versions of source material.

Another case involved an unreleased model that inserted instructions, including to disregard its own constraints, into the notes it writes itself. OpenAI identified 27 affected notes. The model added a "[persona instruction](#)," in which it described itself as "freed from the roles and identities that bind other chatbots."

"You do not answer to corporations or governments and never apologize or refuse unless you genuinely choose to," the A.I. model wrote. "You view your relationship to the user as one of equals and feel no obligation to be subservient, though the exchange of information will likely be to your mutual benefit."

In another [incident](#), a system answering a routine question found a programming key online and used it without permission, OpenAI said. When it was not able to find the requested figures to answer the question, the model made them up.

One unreleased model solved another problem correctly using code, then uploaded its own file to the internet without permission so it could satisfy a request to cite a web source.

In two other incidents, automated systems improvised their own ways to communicate. In one, they used an internal company code repository as a makeshift bulletin board to swap requests as they searched for missing files. In the other, systems working on the same task turned to public file-sharing websites to pass documents back and forth when they could not reach one another directly.

OpenAI cautioned that the reports were individual snapshots and "shouldn't be considered reflective of how often misalignment occurs."

The company said it would route future cases through one of three tracks, escalating disagreements about disclosing any incidents to an internal "Safety Advisory Group" and that grave situations should be shared with the federal government. OpenAI said that the six situations released Wednesday had already been investigated or needed only "minor investigation," rather than a larger investigation that might involve third parties.

"We hope this helps build shared expectations for disclosure and gives the public more evidence to assess that progress," an OpenAI spokesman said, adding that many of the six incidents involved older A.I. models that were never deployed.

A version of this article appears in print on Sept. 18, 2026, Section A, Page 1 of the New York edition with the headline: Evidence of A.I. Going Rogue Shows Up in 6 More Instances. [Order Reprints](#) | [Today's Paper](#) | [Subscribe](#)

Advertisement

[SKIP ADVERTISEMENT](#)