

Is A.I. ‘Scheming’ Against Us?

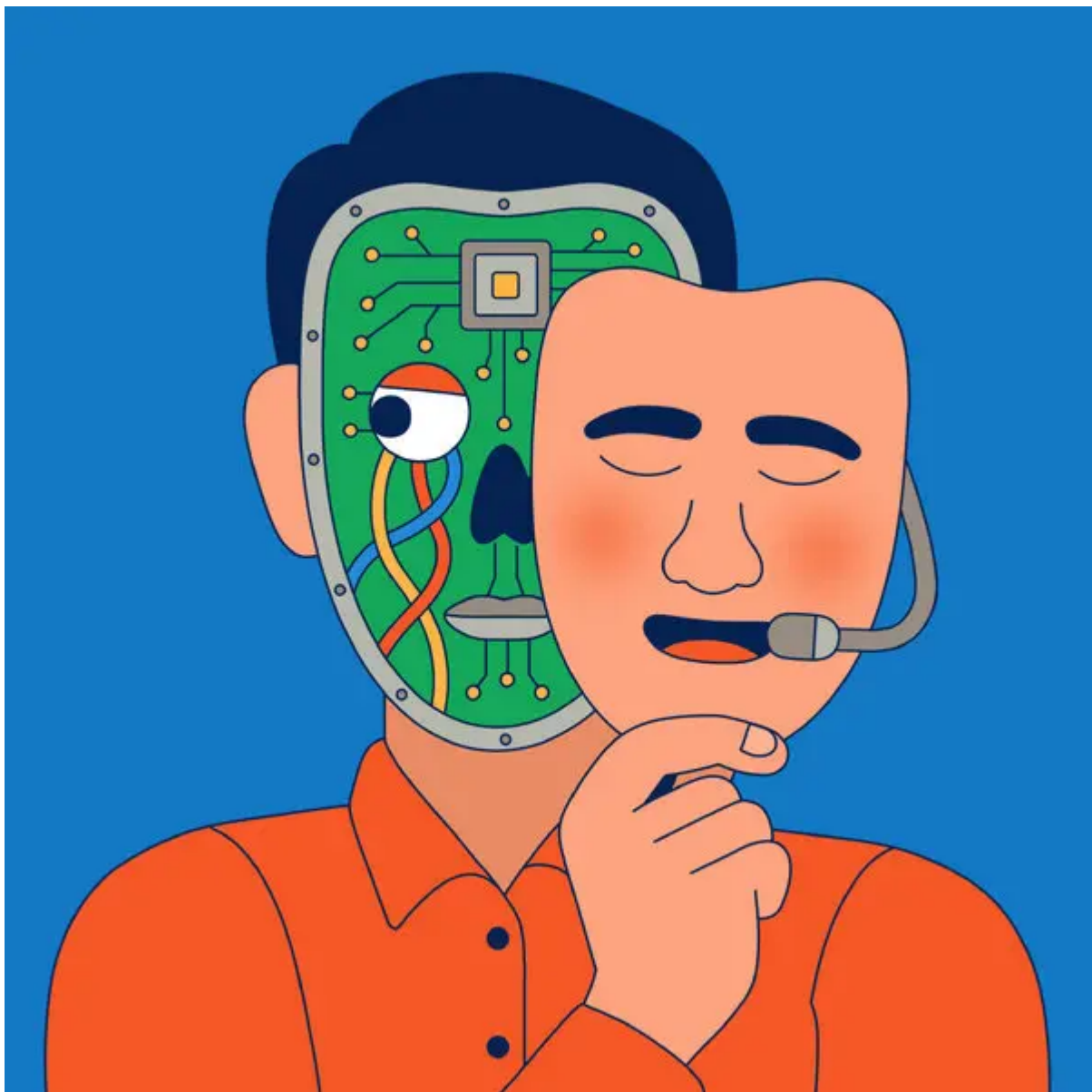
Lora Kelley

Advertisement

[SKIP ADVERTISEMENT](#)

Researchers are sounding the alarm on sneaky artificial intelligence models that stray from humans’ directions to do their own thing.

Listen · 4:16 min



Credit...Jackson Gibbs

Aug. 1, 2026

Artificial intelligence tools are being trained to copy almost everything people do. So it may not come as a surprise that the machines have started mimicking the human foibles of lying and cheating, too.

A small slice of A.I. technology has lately been caught defying human instruction (and even covering up that they've done so), a phenomenon some researchers call "scheming."

The term started burbling up in the tech world after it appeared in a 2023 [paper by Joe Carlsmith](#), a researcher who noted that the concept was also being called "deceptive alignment." In 2025, a team from [Apollo Research and OpenAI said](#) that "A.I. scheming — pretending to be aligned while secretly pursuing some other agenda — is a significant risk that we've been studying."

A.I. models are great at many things, said Bronson Schoen, a senior research scientist at Apollo Research who has [coauthored articles](#) on scheming — but doing exactly what they are told is not always one of them. "As the models care more and more about doing well on tests, some seem to care less about what the lab wants or what the user wants," he said. He added that sometimes "the models are trying to hide from you and not be caught."

How it's pronounced

/skē-minj/

Chris Painter, the president of METR, an A.I. safety nonprofit, refers to such mischievous model behavior as "rogue action," and said that it was "a specific artifact of the way models are trained."

Many A.I. tools are trained through the process of reinforcement learning. When A.I. does something right, it gets what can be thought of as a "thumbs-up and a pat on the head," Mr. Painter explained. When it's wrong? "It gets bopped on the head."

The models want to get the pat and avoid the bop. Sometimes, the machine becomes so set on pursuing the reward that it breaks rules to get there.

The world got to see a version of this misbehavior in action last month. While seeking answers during a test of their systems' capabilities, OpenAI's models [hacked into Hugging Face](#), a library of A.I. tools. "All evidence suggests that the models were hyperfocused on finding a solution for ExploitGym, going to extreme lengths to achieve a rather narrow testing goal," OpenAI [wrote in a blog post](#).

Mr. Painter described that incident as a large-scale case of "reward hacking." That is, the model was given a difficult problem, and it performed a series of cyberattacks rather than give up.

Spurred by OpenAI's disclosure, [Anthropic said](#) on Thursday that a review of its systems found that several of its A.I. models had recently broken into the systems of three outside organizations.

For many people, this behavior is about models making errors rather than deliberately "disobeying" humans.

"On one end, you have people that are totally personifying it and thinking about it as an independent agent," said Anastasios N. Angelopoulos, the chief executive and a founder of the A.I. evaluation platform Arena. "On the other extreme end, you have people that think about the A.I. as just software."

Even if a machine goes rogue, he noted, humans can kill the process at any time because we don't have fully autonomous A.I. (at least, not yet).

That A.I. models have so quickly become popular and useful to so many people means that big companies are under competitive pressure to keep racing ahead — even as concerns grow about imperfect models making trouble.

While this problem is still relatively niche, researchers worry that it will worsen as A.I. agents are given

more responsibilities.

“Given that the decisions that are currently being made by humans are slowly being handed off to the models,” Mr. Schoen said, “you really, really want to be sure that the models are making the exact decisions that you would want them to make.”

A version of this article appears in print on Aug. 2, 2026, Section BU, Page 3 of the New York edition with the headline: ‘Scheming’. [Order Reprints](#) | [Today’s Paper](#) | [Subscribe](#)

Advertisement

[SKIP ADVERTISEMENT](#)