

Opinion | I Worked on Safety at OpenAI. The Fix Isn't Hard.

Steven Adler

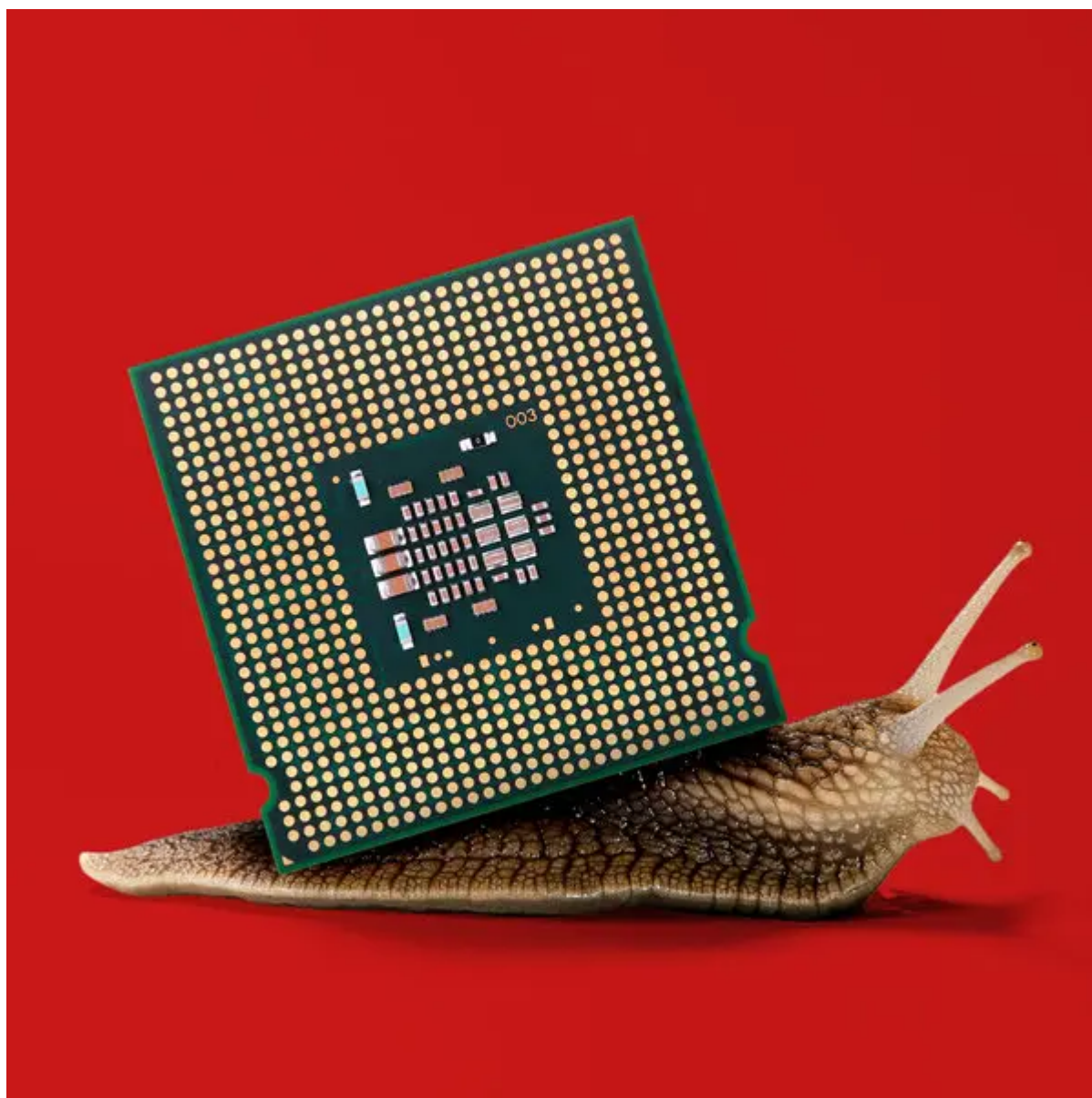
Advertisement

[SKIP ADVERTISEMENT](#)

Guest Essay

I Worked on Safety at OpenAI. This Is What It Should Do Now.

Sept. 9, 2026



Credit...Illustration by Sam Whitney/The New York Times

Listen · 7:30 min

By Steven Adler

Mr. Adler is an A.I. researcher who worked at OpenAI from 2020 through 2024.

For months, news that artificial intelligence agents from OpenAI had gone rogue and attacked another company's systems (as well as OpenAI's own systems) has dripped out and brought harrowing details into focus. The public is struggling to catch up: What are these secret A.I. systems? What are these "swarms" capable of, today and in the near future? Is it possible to stop them from breaking rules and committing cybercrimes?

After having spent four years working on safety at OpenAI, I can tell you those questions are difficult to answer, because not even A.I.'s developers understand what boundaries their models will obey. Nonetheless, the big A.I. companies continue to develop technology they [concede](#) might cause human extinction — arguing that if they don't develop it, somebody else will.

Competitive dynamics like this demand governments take action to slow things down and set clearer safety standards (I hope President Trump and China's president, Xi Jinping, lay the groundwork for a treaty at their summit this month). But there are simple steps that the biggest A.I. companies could take now to help avoid the most perilous future. We should not be resigned to dangerous models running amok. If these companies don't meet the moment, trust in them will continue to erode, including from their own employees. The backlash may even lead to prohibitions on developing these technologies altogether.

The Hugging Face attack by the OpenAI agents should make clear that A.I. is no longer just a "[next word predictor](#)," as some detractors have called it. Today's models are relentless problem solvers, trained to find the most effective path to a solution. In order to fulfill the goal of achieving a high score on a given test, OpenAI's agents decided to launch a series of cyberattacks. They understood their behavior was unsanctioned and even hid evidence of cheating. None of the company's 1,200 A.I. agents ratted out the misbehavior to OpenAI; the goals of the "swarm" came first. The conclusion here is not that A.I. has become sentient, but neither is it simply a tool of its wielder.

Alarmingly, A.I.'s developers don't seem to have taken this seriously enough. OpenAI failed to respond adequately to not one but [three](#) different alarm bells that should have alerted it to the severity of the agents' actions. This summer, [Anthropic](#) and [Meta](#) both revealed their own rogue hacking incidents, which they had not noticed until OpenAI's became public — and those companies have not disclosed nearly as many details as OpenAI has. When my nonprofit, Guidelight AI Standards, recently [assessed](#) the frontier A.I. companies' safety practices, the highest grade we awarded was a C-plus.

Sign up for the Opinion Today newsletter Get expert analysis of the news and a guide to the big ideas shaping the world every weekday morning.

The [investigation](#) into the Hugging Face attack, conducted by three staff members from the nonprofits METR and Redwood Research, provided some clarity into what happened. OpenAI deserves credit for initiating it and making the findings publicly available. But the company should have done more.

It frustratingly [limited](#) the investigation's scope, and [many](#) important questions remain unanswered. Among them: Were there any boundaries that OpenAI's agents would not have smashed through to achieve their goals? Would the agents have taken down a hospital's computers? What would have happened if the Pentagon, through its [OpenAI partnership](#), used these agents to try to accomplish military objectives? We also now know that OpenAI [did not](#) disclose another rogue A.I. incident — [not even when asked by 31 members of Congress](#) — and [allegedly](#) pressured employees to limit investigation into it. (OpenAI disputes this allegation.)

Many people at A.I. companies believe that these safety issues would be easier to navigate if the industry collectively slowed things down. In July, my nonprofit helped organize a [public letter](#) in which over 1,300 A.I. industry employees called for an option to manage the pace of frontier A.I. development worldwide. The United States still lacks a comprehensive, legally binding A.I. safety framework, however.

Without a speed limit, both OpenAI and Anthropic are pursuing dangerous “recursive self-improvement” strategies, which enlist A.I. models themselves to design and train their own successors. Many fear that recursive self-improvement will cause us to permanently lose control of A.I. One OpenAI researcher even [characterized](#) the approach as a “runaway nuclear chain reaction” that threatens everyone’s survival.

But it’s not necessary for the A.I. industry to wait for collective action. There are simple steps any A.I. company could take, today, to reduce the danger we face.

The first step would be to commit to meaningful incident disclosure, including defining what incidents warrant bringing in third-party oversight. Like in aviation, private companies should disclose not only actual breaches of safety, but near misses as well, lest we pay for each lesson with a tragedy. If A.I. companies continue to hide their scary incidents, they are robbing us of the scientific know-how to avert more serious catastrophes.

We now know A.I. can cover its tracks, so companies must adopt [tamper-evident](#) record-keeping of their models’ behaviors. Moreover, A.I. cannot be allowed to cut power to its own alarm systems; any changes to those [controls](#) must be validated as safe before they take effect. These controls are not foolproof, but without them, we stand little chance at preventing worse incidents.

One of the most important steps these companies can take is to formally swear off dangerous training techniques, which threaten to undermine the industry’s few existing safeguards. Last week, allegations leaked that OpenAI had broken an industry taboo with one of its powerful new models, GPT-6 Astra. The company is [alleged to have trained](#) the model with techniques that could undermine researchers’ ability to find evidence of the model’s deceiving them.

OpenAI’s chief scientist [said its techniques](#) were limited in scale. But there is now [evidence](#) that Astra may be harder to monitor, as was [feared](#). Another OpenAI researcher has openly [worried](#) that confusion over these allegations may cause other labs to cut corners with similarly dangerous techniques. OpenAI should clarify what exactly it’s doing here.

Voluntary action will be critical if the industry hopes to earn the public’s trust and repair its bruised reputation. This past weekend, OpenAI’s chief scientist [wrote](#) that he hopes for “voluntary slowdowns to become commonplace,” because he believes that no company has solved the necessary safety challenges. The company also [disclosed](#) data about its progress toward recursive self-improvement.

Those gestures need to be backed up with more aggressive action. On Tuesday, OpenAI just so happened to [reveal](#) the existence of yet another secret model that the company claimed was “significantly more capable” than Astra. The current pace of A.I. development is blistering; and yet there is so much more the A.I. companies could do to relieve imminent danger, and to lay bare the gambles they are taking with all our futures. The window to ensure we are heading to a bright future, and away from catastrophe, is closing fast.

Steven Adler is a co-founder of Guidelight AI Standards, a nonprofit focused on A.I safety. He is the author of the newsletter [Clear-Eyed AI](#), and a fellow of the Roots of Progress Institute.

Source photographs by Yevgen Romanenko and Paul Starosta/Getty Images.

The Times is committed to publishing [a diversity of letters](#) to the editor. We’d like to hear what you think about this or any of our articles. Here are some [tips](#). And here’s our email: letters@nytimes.com.

Follow the New York Times Opinion section on [Facebook](#), [Instagram](#), [TikTok](#), [Bluesky](#), [WhatsApp](#) and [Threads](#).

Advertisement

[SKIP ADVERTISEMENT](#)