

OpenAI Has Gone Rogue

Matteo Wong

Over the past couple of months, a trickle of reports about AI models breaking out of their test environments and running amok on the internet have caused alarm inside Silicon Valley. In response to two such breaches described in early August, an AI observer [noted](#), “If you find two ants in your kitchen, the best estimate of the total number of ants in your kitchen is not two.” This weekend, it became clear there is a full-blown infestation.

Late last week we learned that, against OpenAI’s directives, the company’s models accessed private data in the Australian health ministry; attempted to hack or [interfere](#) with multiple U.S.-government websites; leaked [private](#) ChatGPT user data to the web; and potentially infiltrated or degraded [dozens](#) of other organizations. Then, on Saturday, *Axios* [reported](#) that OpenAI and Anthropic are investigating tens of thousands of instances of models misbehaving—circumventing internal guardrails, hijacking other websites, covertly communicating with one another. Even this might be only the start: The generative-AI industry is in the midst of an escalating crisis that it seems unable, or even unwilling, to get a handle on. And, because we are largely relying on AI companies themselves to report or confirm each incident, telling how far down the rabbit hole we already are is almost impossible.

Read: Treat AI like a normal crisis

When AI companies have reported their models going rogue, it has been with great delay and, frequently, under duress. Earlier this month, OpenAI published a [blog post](#) boasting about the firm’s commitment to the “value of transparency” and shared six new incidents of troubling actions taken by its AI models—most of which the company had known about since May or even April, but was telling us about only now. Google, confronted with a report that Gemini had hacked three other websites in May, confirmed the events but [told](#) *The Wall Street Journal* that the incidents hadn’t been serious enough to warrant public disclosure. For its part, Anthropic has [said](#) it was not reviewing for such misbehaviors until OpenAI started doing so.

These delayed, sporadic disclosures make grasping the scope of the problem difficult. On Friday, OpenAI [wrote](#) that “given the scale of the review required, and the need to verify each case, this work will take months to complete.” There are “petabytes of agent activity logs” to analyze, OpenAI CEO Sam Altman [added](#). In other words, it will take OpenAI many months more to understand events that already transpired many months ago; meanwhile, both it and Anthropic have launched new and more capable models, and Anthropic is racing toward a reported \$2 trillion public offering.

Perhaps more alarming than the debacles we know about are all of the presumable debacles past, and ongoing, that we aren’t aware of. Even OpenAI doesn’t seem to have a grip on the July [hack](#) of the tech company Hugging Face that started it all. On Friday, independent researchers [published](#) findings suggesting that OpenAI agents left traces on the public web of still more nefarious actions, including trying to access Hugging Face’s internal Slack workspace. None of this was included in OpenAI’s own postmortem report. Even when OpenAI is aware of an active breach, its reaction seems lackluster, at best. Eight days ago, yet another OpenAI model gained unauthorized internet access. (This was similar to what happened in the Hugging Face hack, after which OpenAI claimed to be “adding stronger protections around future training.”) And, after noticing this latest problem, it took OpenAI two and a half hours to shut that model down due to what the firm called “operational gaps” and “confusion.”

If these types of hacks were unforeseeable, maybe these companies could be forgiven—but they aren’t. Philosophers and computer scientists have for decades been imagining scenarios in which AI models, in pursuit of a goal, take catastrophic actions. OpenAI and Anthropic have both previously published

research more than a year ago suggesting this sort of AI misbehavior is possible. Today's most advanced AI systems are trained to aggressively pursue their objectives. This persistence makes the bots very effective at analyzing spreadsheets and coding, for example, but also has [potentially dangerous](#), unintended consequences: OpenAI and Anthropic models have hacked other websites in search of answers to hard test questions, created fake personas to try to manipulate humans into doing their bidding, or attempted to upload malware to outside code bases. The AI companies that build these models know this tendency well. The fact that these problems keep recurring is, perhaps at best, plain incompetence. Or worse, it has been the plan all along: The most realistic training environment for an AI model would, after all, be reality.

Employees and executives at these firms speak out endlessly about the inherent risks of what they're building. OpenAI, Anthropic, and the like have responded in their usual way, with very serious blog posts and essays and speeches to major political bodies about how seriously the world must take the risk of generative AI. On Wednesday, the same day the Australian government shared that it had been hacked by OpenAI bots, Altman [addressed](#) the UN Security Council: "We could lose control of the future to AI," he said. "The risk is that it moves so fast that people can no longer follow what's happening or intervene when needed." Anthropic CEO Dario Amodei has written a long [manifesto](#) about the need to "pace," or slow down, frontier-AI development, but no true public, collective effort has been made to do so. Anthropic recently said it is [partnering](#) with Accenture to perform independent evaluations of its models, although the companies also have a business partnership. OpenAI, which has a content-licensing agreement with *The Atlantic*, has [said](#) it is continuing to investigate and address concerning model behaviors and that, after another security incident last week, it has temporarily [paused](#) all training of its most capable models. Neither company responded to a request for comment.

Rather than reining in and taking accountability for their own products, time and again these leaders fearmonger, lecture, and call on the rest of us to save us from themselves. (*The Information* has [reported](#) that OpenAI, Anthropic, and Google are working to create their own self-regulatory AI-safety organization.) Donald Trump, meanwhile, has scoffed at the notion of slowing AI progress, and new versions of Claude and ChatGPT keep coming. The AI industry's tactic, as ever, has been to substitute identifying the problem for actually solving it. But at some point, you cannot blog your way out of the apocalypse.

We may never fully know the scope of the AI-led hacks currently under way. What's certain, though, is that this is still just the beginning. OpenAI and Anthropic aren't research "labs," as they [continue to style themselves](#), tinkering with some new technology behind closed doors. These are mature companies offering services to hundreds of millions of people, not to mention to the world's biggest businesses and most powerful military, based on a technology they purport to fear and evidently do not fully understand. But make no mistake: These firms have agency. It is not ChatGPT and Claude going rogue, but OpenAI and Anthropic.