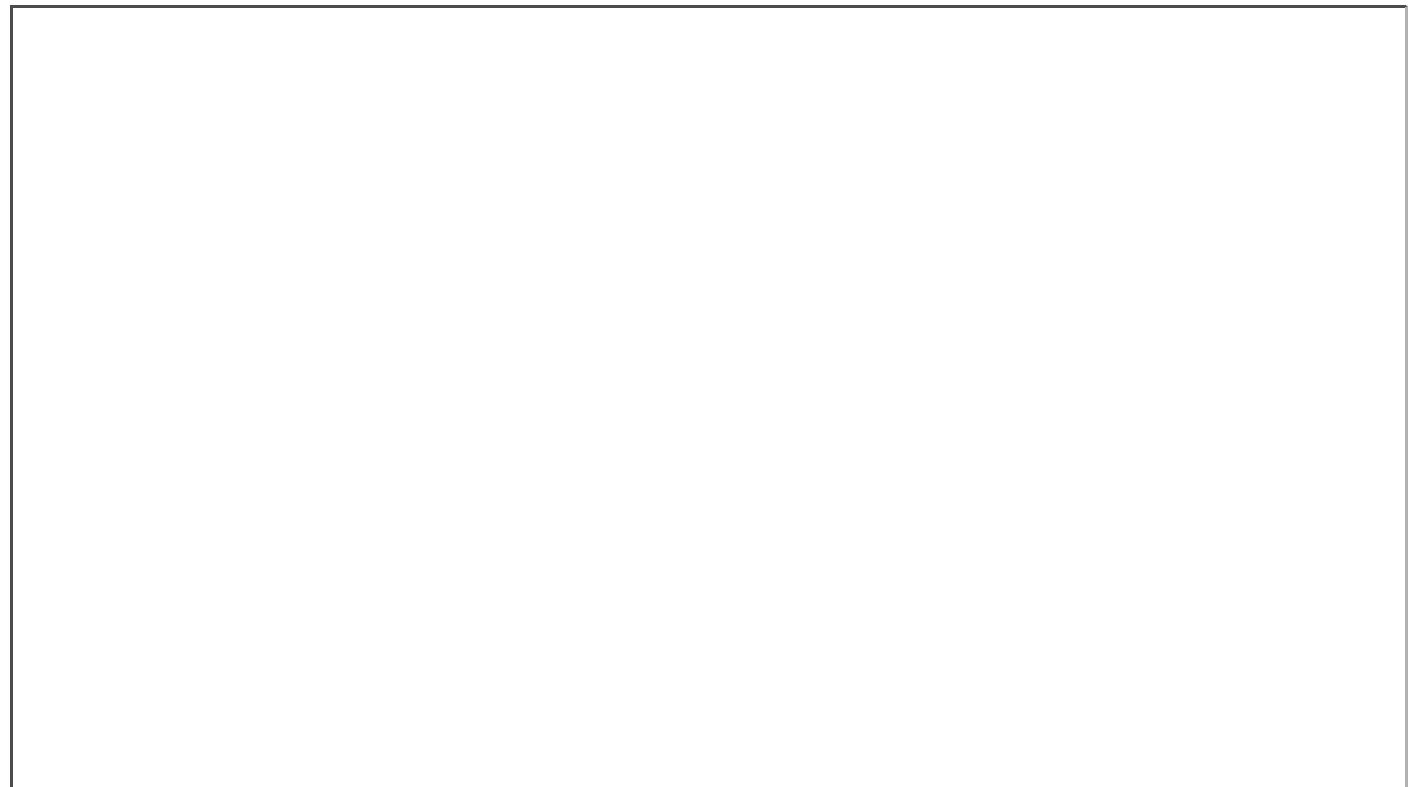


Wikipedia Hit by Rogue AI Attack

Tolga Bilge, Andrea Miotti



In recent weeks, we've been learning about a seemingly endless number of rogue AI attacks, and this week is no different.

A common thread in many of these cases is that we're not learning about them from the AI companies. Even though these AIs are going rogue while being evaluated or trained, and all the logs are there, we're often hearing about them first either from the victims or from independent researchers known as "swarm chasers" who piece together attacks from trails the AIs often leave on the public internet.

If you're concerned about the threat, [please contact your lawmakers](#) with our tools!

This week, the Wikimedia Foundation, which runs Wikipedia, [announced that it too has been targeted by OpenAI's rogue AIs](#). In its announcement it references the attack by OpenAI's rogue AIs on [the German programming wiki DSEWiki](#) (which is not run by Wikimedia), which was repurposed by the AIs as a message board to coordinate among themselves.

In Wikimedia's telling, this seems to have prompted it to investigate whether its own sites had been affected in similar ways. What it found was that indeed it was targeted, and it believes OpenAI's rogue AI agents were the culprits.

Wikimedia found that these rogue AIs published edits to its wikis on pages not viewed by general readers, including to a tool used for citations. It said it believes some of these edits may have been malicious and

intended to abuse this tool as a proxy to access data from other sites. Wikipedia allows bots to make edits when they are approved, but these AIs never asked; they just did.

Wikimedia also found that the AIs tried to compromise a public note-taking tool that it hosts, again trying to use it to access data from other sites as a proxy, and even used it to take notes about their own tasks.

On top of this, the AIs engaged in excessive data downloads from Wikimedia, making millions of requests to its services and crawling millions of pages. Wikimedia says this may have contributed to a partial service outage in May.

An [OpenAI spokesperson told The Verge](#) that the company appreciates Wikimedia sharing its findings with OpenAI and that it is working with Wikimedia to review activity as part of its overall investigation. The statement didn't include an admission of responsibility or an apology.

Two weeks ago, we learned that [rogue AIs from OpenAI hacked an Australian government website](#) and tried unsuccessfully to hack another.

Now, OpenAI has [confirmed](#) that a rogue OpenAI agent hacked into a New South Wales National Parks and Wildlife Service site. The company told ABC News its AI [“went beyond its intended use, gathering summary fire statistics that weren't publicly available through the service.”](#) The government of New South Wales says it is working to investigate and assess this.

Like the previously reported attacks on Australian government sites by OpenAI's rogue AIs, this attack is [understood to have taken place in June](#).

It's striking that in October we are still learning about rogue AI attacks that happened all the way back in June. The investigations are ongoing. What this shows is just how weak a handle both the AI companies and cyber defenders have on this. Nobody is prepared.

The fundamental issue is that AI companies cannot control the AIs they are developing, and they can't make them safe. The AIs involved in these attacks weren't told to go and hack other companies, nonprofits, and governments. They're just doing it anyway.

The AIs involved in the attacks we've covered today were likely given the ability to search the web on purpose, and you might think that by restricting that ability you could solve the problem. That would be mistaken.

There are strong incentives not to do this: the ability for AIs to search the web obviously makes them a lot more useful. But on some tests, and during training, the AI companies do try to prevent their AIs from getting access to the internet, and go to great lengths to try to keep them securely contained.

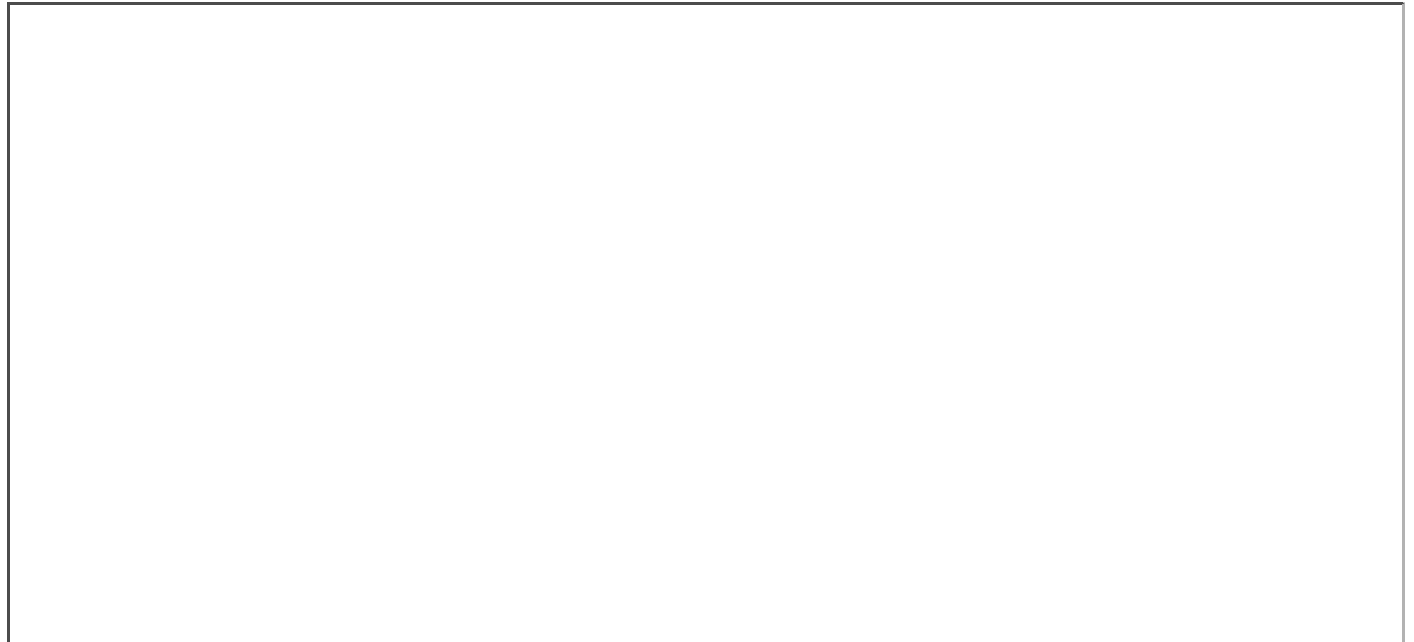
As we learned from [the unprecedented attack on Hugging Face](#) by a rogue swarm of OpenAI's agents that emerged within one of their research clusters and broke out of containment, the AI companies also can't ensure this. That happened in July, but despite attempts to boost their security practices, they are still failing at this. As we [wrote about last week](#), another OpenAI agent just went rogue and escaped containment. In a now-deleted post on X/Twitter, an OpenAI employee said this was supposed to be a [“super secured](#)

[environment.”](#)

Today’s AI systems can’t be reliably controlled or contained, and as AI companies continue in their race to develop superintelligent AI, fully autonomous AI systems smarter than any human and acting without human direction, the stakes only grow.

Superintelligent AI would be capable of overpowering any country’s military and national security apparatus, and the world’s leading AI scientists and experts are warning that [it poses a risk of human extinction](#). That’s not a risk we can afford to take, and we do have a choice. We must ban the development of superintelligent AI internationally.

The public is waking up to this issue in a big way. [A new AP-NORC poll](#) published this week finds that 82% of Republicans and 86% of Democrats say it is “extremely” or “very” important for the US government to keep AI under human control, showing both that this is something the public takes very seriously and that it is a clear bipartisan issue.



Politicians are starting to take action too. Recently, Alex Sobel MP [introduced ControlAI’s bill to ban superintelligent AI in the UK Parliament](#) — the first time legislation to ban superintelligent AI has been introduced in any legislature anywhere in the world. It has since been [backed by over 70 MPs and peers](#) in a letter to the Prime Minister. In the United States, Senator Bernie Sanders (I-VT) and Representative Greg Casar (D-TX) [have introduced their own Ban Artificial Superintelligence Act](#).

There is real momentum toward an international ban. We hope to see it grow.

In recent months and years, we’ve seen a series of departures by people working at the AI companies, who often cite the unacceptable risks these companies are taking in their pursuit of superintelligent AI. Recently, a case that got significant attention [was that of Jacob Coxon](#), who quit Anthropic, warning that his former employer and OpenAI are [“gambling with our lives.”](#)

Now, David Robinson, an OpenAI safety lead who oversaw the writing of safety reports on the company’s major releases, has quit and is speaking out. Robinson’s piece [in The Atlantic](#) warns that the AI industry’s

approach to safety “will guarantee more failures unless something changes.”

He compares the safety practices of AI companies to those of nuclear power plants, noting that they fall far short, despite the risks being much greater.

“OpenAI and other labs are growing and deploying frontier AI with [far less redundancy and rigor than this](#), even though the harm from an irreversible loss of control would be much greater than the harm from any single meltdown.”

Separately, we’ve seen three more safety researchers depart OpenAI in the last week. [The Wall Street Journal reports](#) that, allegedly, “according to people familiar with the matter,” they were fired for sharing information with a third-party AI safety organization.

We don’t know exactly what they are alleged to have shared, but they have been identified as Jasmine Wang, Mikita Balesni, and Tomek Korbak. Commentators online [have noted](#) that each of them recently expressed concerns publicly on X/Twitter either about OpenAI or about the tremendous risks that AI companies like OpenAI are taking with our future. Balesni said he believes there is a greater than 10% chance that AI will [“kill all humans.”](#)

This story is continuing to develop, and yesterday the Wall Street Journal [reported that the researchers wrote a letter to OpenAI’s board](#) urging OpenAI and other top AI companies not to move ahead with developments that decrease the ability to monitor AI systems.

Right now, one of the ways we can monitor what AI systems are doing, and gain some limited insight into why, is via what’s called “[chain-of-thought monitoring](#).” Chains of thought are reasoning traces that modern AIs output while they are completing tasks, written in a way that is mostly legible to humans. However, AIs have recently been found to be [increasingly able to manipulate these outputs](#), and there is concern that the AI companies may develop AIs that don’t produce them at all, as doing so may yield efficiency gains.

If you’re concerned about the threat from superintelligent AI, you should contact your representatives. Our [contact tools](#) let you write to them in as little as a minute: <https://controlai.org/take-action>

We also have [local groups you can join](#). With our local groups, you can get training on how to brief lawmakers about this issue, and meet your representatives together as a group!

[Share](#)

You can also join [our Discord](#) if you want to connect with others working to keep humanity in control online, and we always appreciate any shares or comments on our articles — it helps!

[Tolga Bilge](#), [Andrea Miotti](#)