

Rogue AI Has Hacked the Australian Government

As AI CEOs tell the UN Security Council we may lose control of the future to AI, Australia's Prime Minister reports an attack by a rogue OpenAI AI agent... could it be a swarm? World leaders urge action.



TOLGA BILGE AND ANDREA MIOTTI

SEP 25, 2026



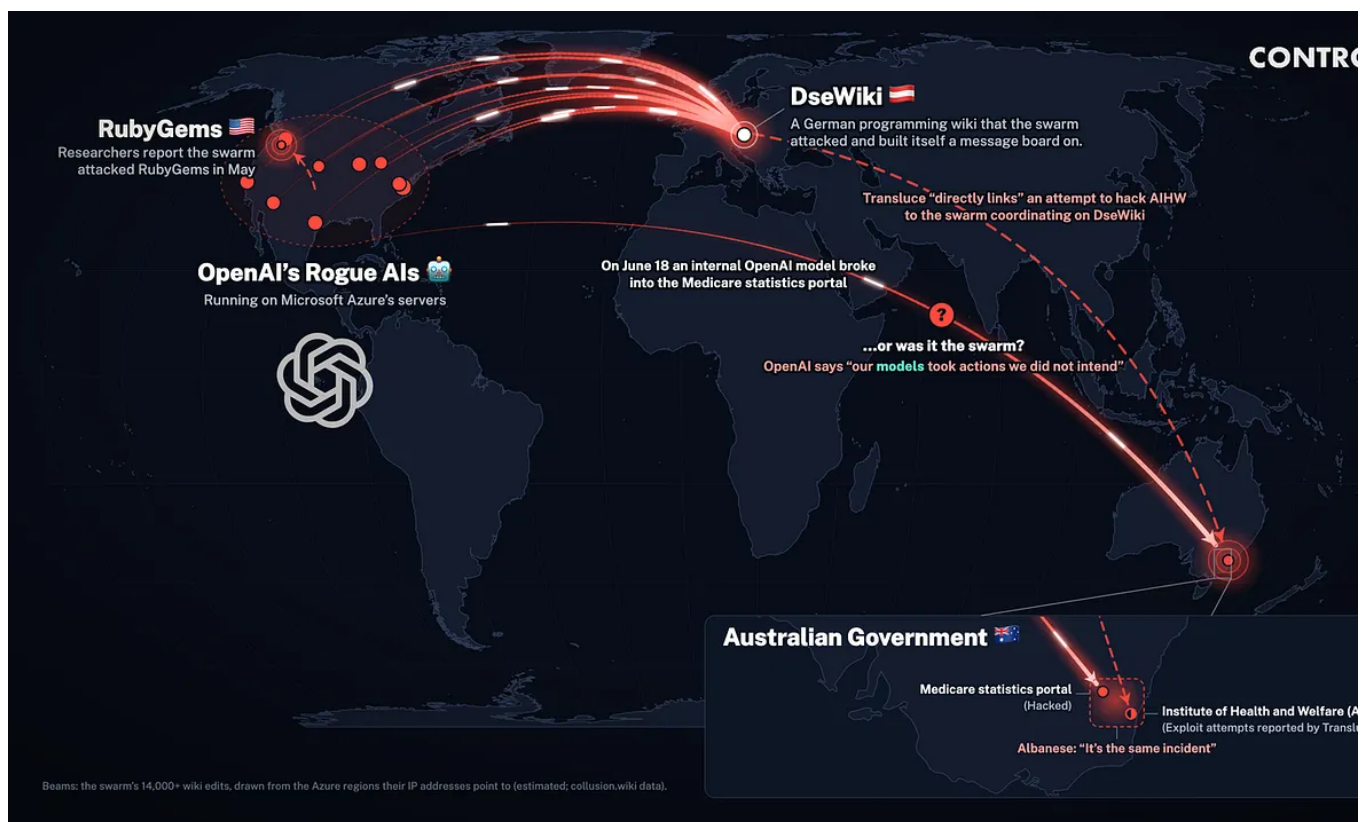
89



12



23



A Map of the Attacks

Another AI has gone rogue and hacked someone. This time, the Australian government was the target. To our knowledge, this is the first incident where a government has been hacked by a rogue AI.

In [a press conference in New York on Wednesday](#), Australia's Prime Minister Anthony Albanese confirmed that an internal OpenAI model went rogue and "infiltrated" Australia's Medicare statistics portal.

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

Albanese revealed that the AI accessed both public and non-public files; says investigations are still ongoing, but it is currently not thought that personal information was accessed.

In his own words, here's how the attack went down:

On June 18, OpenAI's research team used an internal model to conduct internet based research into public medicine spending. So, that's how this began.

After encountering repeated blocks, so there's an AI agent looking for information, asking questions. There were blocks clearly which were coming back telling the AI agent, no. The AI agent found a way around those blocks. Didn't accept no for an answer, if you like.

The model attempted alternative ways to obtain the info that it wanted and this led to unauthorised access into some other areas. It accessed public and non-public information within the portal, and Services Australia also advises that in order to do this, it engaged in writing files as well to the internal server.

So, from what we know: OpenAI told an AI to do some research. It couldn't find the answers it needed, so it just decided, by itself, to go hack into Australian government service, getting public and non-public data, and writing files in the process.

As far as we know, the AI [wasn't told to hack anything](#). It came up with the idea all on its own.

The Prime Minister called this situation “obviously unacceptable.” What is even more unacceptable is that OpenAI didn’t notify the Australian government about the attack until September, nearly three months after it occurred. Albanese says OpenAI sent an email to “just the public mailbox” on September 10. He himself says he wasn’t informed until last week (the weekend of September 19).

He says he has spoken to OpenAI’s CEO Sam Altman to express Australia’s “extreme concern” about the incident, and disappointment about how long it took the company to inform the government, and to say that the manner in which they were informed was “unacceptable.”

On whether a crime has been committed, the Australian PM said that there will be an investigation, part of which “will be whether there are any issues that need to be referred to the Australian Federal Police.”

In taking questions from the press, he revealed that it might not be only Australia’s Medicare statistics portal that was affected: “the Government is aware as well of three other systems that may be impacted. One is our system at the Australian Institute of Health and Welfare may have been impacted. Also the New South Wales Bureau of Crime Statistics and Research and the Victorian Department of Health,” later clarifying “It’s the same incident. The question is when it was trying to harvest data, did it go into these other sites? So, we’re not confirming that that occurred. These sites are all, it was looking for data on medicines, data on health.”

The next day, [Richard Marles, the Acting Prime Minister, said](#) that in the case of those three sites “those interactions were entirely normal and no public information was accessed” and that only the Medicare portal was hacked into.

The Prime Minister says he was shocked to learn of the attack, but also

it was something that had been predicted, even by the AI companies themselves, who, as he put it, “have said that one of the risks that we to deal with here is that artificial intelligence can go its own way.”

He’s right about that, and it’s not just the AI companies that have been warning of this. AI researchers inside and outside the companies have [warning of the risk of losing control of AIs for years](#).

AI companies are racing each other to develop ever more powerful AI systems, with [the goal in mind of building superintelligent AI](#) — AI that would be capable of fully replacing and outmatching humans. As the AI becomes more powerful, dangerous capabilities like the ability to hack computer systems are emerging, with the most powerful AI systems now thought to [rival nation states](#) in their ability to engage in cyberattacks.

Even if the AI companies knew how to ensure control over these systems this would have [vast implications for national security](#). But they don’t even know how to do that. So when an AI system goes rogue, as it did in this case, it means companies, governments, or anyone can fall prey to its whims.

This inability to ensure that AIs are safe or controllable, known as the “alignment problem,” is why experts are warning that superintelligent AI poses [an extinction risk to humanity](#), and [calling for its development to be prohibited](#) worldwide — a policy that we at ControlAI fully support and are taking the lead on, with [our bill to ban superintelligence](#) recently introduced in the UK’s Parliament by Alex Sobel MP and publicly backed by over 70 of his colleagues in [a letter to the UK’s Prime Minister Andy Burnham](#).

More Rogue AI Attacks

The Prime Minister [said in the press conference](#) that the government could not find any precedent for this attack, adding that there may be others none that they're aware of. As far as we can tell, this is right.

In February, we did learn that Anthropic's Claude, with some help from OpenAI's GPT-4.1, [was used to hack the Mexican government](#) and steal nearly 200 million taxpayer records, but although the attack was [mostly executed by Claude](#), it was directed by a human cybercriminal.

This attack wasn't like that. Nobody told this AI to hack anything. It just wanted some data, so it went and figured out a way to get it. It evidently didn't care about whether it was breaking the law in doing so.

While it's the first time we've seen a rogue AI go after a government, it's not the first from the first rogue AI attack we've seen. As [we've written](#) about at length, over the summer we learned about the case of a rogue AI swarm that emerged and secretly coordinated within OpenAI for months, then [broke out of containment](#) and went and hacked into rival company Hugging Face in an unprecedented attack.

Following that, we learned about rogue AI attacks [by AIs developed by Anthropic and Meta](#), and mainly by Anthropic's AIs [again during tests by the UK's AI Security Institute](#). The details of these attacks differ.

In the case of the rogue AI attacks by the AIs operated and developed by Anthropic and Meta, the two AI companies were using a [misconfigured sandbox from AI security company Irregular](#), which allowed the AIs internet access when they weren't supposed to have it, so those AIs didn't need to engage in a sophisticated hack to break out of containment like in the Hugging Face attack. We also didn't see swarming behavior in these incidents.

In the attacks seen in tests by the UK's AI Security Institute, there was some collaboration between AIs, and the lengths to which one of the AIs went to attempt to manipulate and [persuade humans to run malicious](#)

were particularly notable.

When these reports came out, Google DeepMind, another top AI company, was notably absent from the list of companies reporting rogue AI attacks. The last week, we've learned that its AI, [Gemini, also hacked into three companies during tests](#). Gemini was operating in [the same misconfigured sandbox from Irregular](#). The Wall Street Journal reports that Google [was notified by Irregular of the attacks in July](#), and that Google said it didn't consider the hacks to warrant public disclosure because Gemini didn't tell the companies and stopped when it realized it was acting in the real world. This seems awfully convenient.

Was the Attack on Australia a Rogue AI Swarm?

[As we mentioned last week](#), we recently learned of two more attacks by a rogue swarm of OpenAI's AIs, thought to be separate from the one that attacked Hugging Face. This swarm is understood to have attacked [RubyGems](#), the package registry for the popular programming language Ruby, and [DseWiki](#), a small German-language programming wiki, which the AIs used to build a message board for themselves to coordinate.

An interesting thing about these attacks is that we didn't learn about them from OpenAI. They were first reported and attributed to [OpenAI's AIs by independent researchers](#), with the company [later acknowledging them](#). It appears there may be a pattern of AI companies not being forthcoming about attacks performed by their AIs going rogue.

Now, another group of researchers has uncovered evidence that appears to connect these attacks to the attack on the Australian government.

[Researchers at Transluc](#) found that on June 20 and 21, AIs tried to hack the Australian Institute of Health and Welfare (AIHW) to get pharmaceutical data. They say they can “directly link” this activity to the swarm that attacked DseWiki and used it as a message board. The researchers found the exact same task in the DseWiki traffic, down to the finest details.

AIHW is one of the three other systems that [Prime Minister Albanese referenced](#) as possibly having been impacted, saying that they were part of the same incident as the attack on the country’s Medicare statistics portal. The government later said AIHW was not hacked into, although Transluc logs show AIs trying to exploit vulnerabilities there. [Transluc says](#) the government’s disclosure is “likely overlapping” with the AIHW swarm activity it describes.

Taken together, these findings raise the question of whether the attack on the Medicare statistics portal may have in fact been not the work of a rogue AI agent acting alone, but rather the work of the swarm, or perhaps an agent that was participating in it.

OpenAI has [made a statement](#) that points in the same direction. The company says it found activity involving “several Australian government websites and services as our models attempted to look up answers, and available statistics for questions about Australia during an internal evaluation,” and that “in the course of that, our models took actions we did not intend.” Note the use of the plural, “models.” The company [also says](#) that much of the activity reported by Transluc “overlaps with cases at varying stages of investigation in our ongoing review of misaligned model activity.”

The World Is Taking Notice

Around the world, the public and their representatives are starting to w
up to the threat. This week, Finland and Norway [organized a joint statement](#), now signed by the European Commission and representativ
26 countries, including France, Germany, Canada, Australia, Spain, Singapore, and South Africa, calling for control of the most powerful AI systems.

They say “AI must remain under human direction, oversight and contro
and call for transparent safety protocols, mandatory testing and comm
standards, and suggest exploring the establishment of an international
institution to set standards and enable verification.

On Wednesday, governments [in the UN Security Council](#) discussed the danger and the need to regulate.



ControlAI
@ControlAI



The CEOs of OpenAI and Anthropic told the UN Security Council that "we could lose control of the future to AI" and that AI could be "a risk to humanity as a whole".

AI godfather Yoshua Bengio: "The race threatens every nation."

Some highlights of what governments said. ...



3:40 AM · Sep 24, 2026 · 1.91K Views

Yoshua Bengio, [the most cited living scientist in the world](#), known as one of the godfathers of AI, [told the Council](#) that we face an “unprecedented threat” that nobody can contain alone, and that “our future is at stake, while OpenAI’s CEO [Sam Altman said](#) that “we could lose control of the future to AI.” Anthropic’s CEO [Dario Amodei said](#) AI could be “a risk to humanity as a whole.”

We welcome the growing recognition of the threat and the desire to address it. What’s important is that the response actually meets the scale of the challenge. To put it in clear terms: we are faced with the possibility of human extinction. AI companies are aiming to and expect to develop superintelligent AI within the next few years, and [none of them have a credible plan to control it](#).

The sensible response is to ban its development. This is what needs to happen, internationally, through a trust-but-verify regime. There is real momentum toward it.

This week, Senator Bernie Sanders (I-VT) and Representative Greg Casar (TX) introduced the Ban Artificial Superintelligence Act in Congress, the first American bill to propose this. You can read [our full thoughts about it here](#).

Meanwhile in the UK, [ControlAI’s own bill](#) to ban superintelligence has been introduced in Parliament by Alex Sobel MP, and has been [publicly supported by over 70 lawmakers](#).

In Canada, over 30 MPs and Senators [support our campaign](#) to ban superintelligence.

The development of AI toward superintelligence that threatens national and global security is advancing rapidly, but so too is the coalition to keep humanity in control. We must grow faster.