September 28, 2026 Company Safety

How we will do better for Australia

Listen to article 8:58 Share

In June, during internal training and evaluation our models accessed Australian government websites in ways they were not authorised to. We also should have handled our response better. We are sorry and working to do better in the future.

In this post, we are setting out what we know, what we have changed, and what we will do to rebuild trust with the Australian people. This is a new kind of cyber incident which represents an emerging global challenge. One of the ways we intend to take accountability for the situation is to be intentional in working with Australia to help develop practical approaches to how AI developers and governments identify, disclose, and respond to AI cyber behaviour, whether malicious or unintentional.

When we became aware and how we responded to this incident

After the Hugging Face incident in July, we began reviewing earlier training and evaluation activity to identify other affected organisations. In mid-August, that review identified activity affecting the Australian government websites below.

Here's what we know based on the evidence:

- **Services Australia:** An OpenAI model discovered a way to gain non-public access to the service, and ran commands, retrieved internal files, credentials and aggregate statistics, and wrote files. However, individual patient or client records were not accessed. We cover this incident in more detail below.
- **NSW Bureau of Crime Statistics and Research (BOCSAR):** An OpenAI model accessed BOCSAR's public Crime Mapping Tool to research public crime statistics (we explain further below why models carry out various information research tasks). The model made API and website metadata requests via the public

configuration, operational jobs and logs, and website metadata. Crime records of individuals were not accessed.

- **Victorian Department of Health:** OpenAI agents discovered an exposed access key to query the Victorian Agency for Health Information's reporting system and retrieve reporting configuration and aggregate survey statistics. The extent to which this information should have been accessible is unclear, and depends on VAHI's access policies. Individual medical records or identifiable survey responses were not accessed.
- **Australian Institute of Health and Welfare:** OpenAI agents retrieved aggregate statistics using third-party browsing and download services, including from AIHW's website, and queried chart data directly. Separate attempts to bypass access controls were unsuccessful. The downloaded material appears to have been publicly available. There was no system compromise. Individual medical records were not accessed.

We launched investigations into these activities as soon as we became aware in mid-August. We notified Services Australia and the Victorian Department of Health on 10 September and the NSW Bureau of Crime Statistics and Research on 18 September. The activity related to the Australian Institute of Health and Welfare did not meet our disclosure thresholds because the way it was accessed seemed consistent with public access, but we notified it on 24 September to share our findings and offer a briefing.

Our aim was to give affected agencies a detailed account once our investigation was complete. However, we should have shared preliminary findings sooner and kept Australian agencies updated as more facts emerged.

Since then we've worked closely with Australian government agencies to share what we've learned to date. If we identify any additional affected agencies, we will notify them promptly and directly with the information available and provide updates as further facts emerge.

What happened with Services Australia Medicare Statistics Reporting Service

During internal training and evaluation in June, we were running an experimental, internal-only OpenAI model that was not intended for public release and without the full set of safeguards used in our publicly available products. In the course of this training and evaluation, it accessed Services Australia's Medicare Statistics Reporting Service. Our review to date has found no evidence that anyone's medical records were accessed.

them tasks drawn from a broad collection of research questions spanning many subjects, reflecting the kinds of detailed questions users might ask. This trains a model to find, interpret and analyse publicly available information so the model can be more useful to people. Our models are supposed to answer these questions using publicly published statistics.

In this case, one of the tasks assigned to the model was to research government spending per person on medicines for skin conditions in Victorian communities. The model had difficulty obtaining that information, and it took actions that we had not authorised it to take. In the course of looking for this information at Services Australia's Medicare Statistics Reporting Service, it discovered a way to gain non-public access to the service. It then used this access to review technical system information and source code related to the service—all still with the objective of trying to find the information it was originally looking for. We did not intend for this activity to occur, and the access to the service and follow-on activity should not have happened.

What we are changing

Following the Hugging Face incident, we strengthened our research safeguards, including additional network restrictions and expanded monitoring. We implemented controls to block live internet access in these research environments, with web access served through cached content. As an additional layer of security, our current monitoring systems would have detected this activity and paged our team for urgent human review. For example, when a model gained live internet access during a recent training run, our monitoring detected the activity and paged a human reviewer, and we stopped the run. We continue to test these protections and address gaps. Hugging Face remains the most severe incident we have observed.

People want to know AI is being developed safely, and that starts with what companies like ours do ourselves. We recently shared that we've paused training and evaluation involving tool use for our most capable models and will resume training them only when we are confident that we have additional safeguards in place, which we are working on now.

As AI systems broadly grow more capable, we also see a narrowing window to help organizations find and fix weaknesses. This takes collective action working with defenders worldwide.

We have joined organizations across technology, cybersecurity and critical infrastructure in a call for collective action on cyber defence. That call starts with our own responsibilities including stronger safeguards, timely disclosure and practical support for affected

share what works.

In Australia, we are committing resources and expertise to support affected agencies and help defenders put these principles into practice, including:

- **Dedicated support for affected agencies.** We will commit the resources needed to help affected agencies understand what happened and assess the impact. This includes sharing relevant technical findings and arranging engagement with our response teams through appropriate information-sharing arrangements.
- **Funding and support to strengthen cyber defences.** We will support Australian governments and industry through credits from our $1 billion Daybreak for Frontline Defenders fund and technical assistance to strengthen cyber defences across critical infrastructure and other sensitive environments. Building on our engagement with governments and critical infrastructure operators, this work will help organisations better understand, detect and respond to risks from increasingly capable AI agents.
- **An Australian taskforce.** We will establish a taskforce with independent Australian expertise to develop practical policy recommendations for managing risks from increasingly capable AI agents. Drawing on lessons from these incidents, it will focus on improving notification processes, strengthening coordination between AI developers and government, and identifying measures to better protect government systems. The taskforce's recommendations will inform OpenAI's approach and support the Australian governments' work on AI safety and cybersecurity. The taskforce, which is expected to complete its work by the end of the year, will also recommend practical steps AI companies can take to reduce the risk of similar incidents.

Rebuilding trust with Australians

Australia's governments, industries, and citizens are and have been invaluable partners to OpenAI. We do not take this for granted, and we intend to make this right.

OpenAI's Chief Strategy Officer, Jason Kwon, will fly in from OpenAI's US headquarters to appear at the Joint Select Committee on Artificial Intelligence in Sydney on Tuesday 6 October. He will answer questions about what we know, how we responded, what steps we have taken, and how we will do better going forward.

We will continue sharing verified findings with affected agencies and relevant governments. We will publish updates on our ongoing review and progress against these commitments. We know we have a lot of work ahead of us to rebuild trust, and that we are accountable for

2026

Author

OpenAI

Keep reading

View all



DevDay 2026 Recap

Company Sep 29, 2026



Introducing dots

Product Sep 29, 2026



Towards safety cases for frontier AI training

Safety Sep 28, 2026

OpenAI © 2015–2026 Your privacy choices