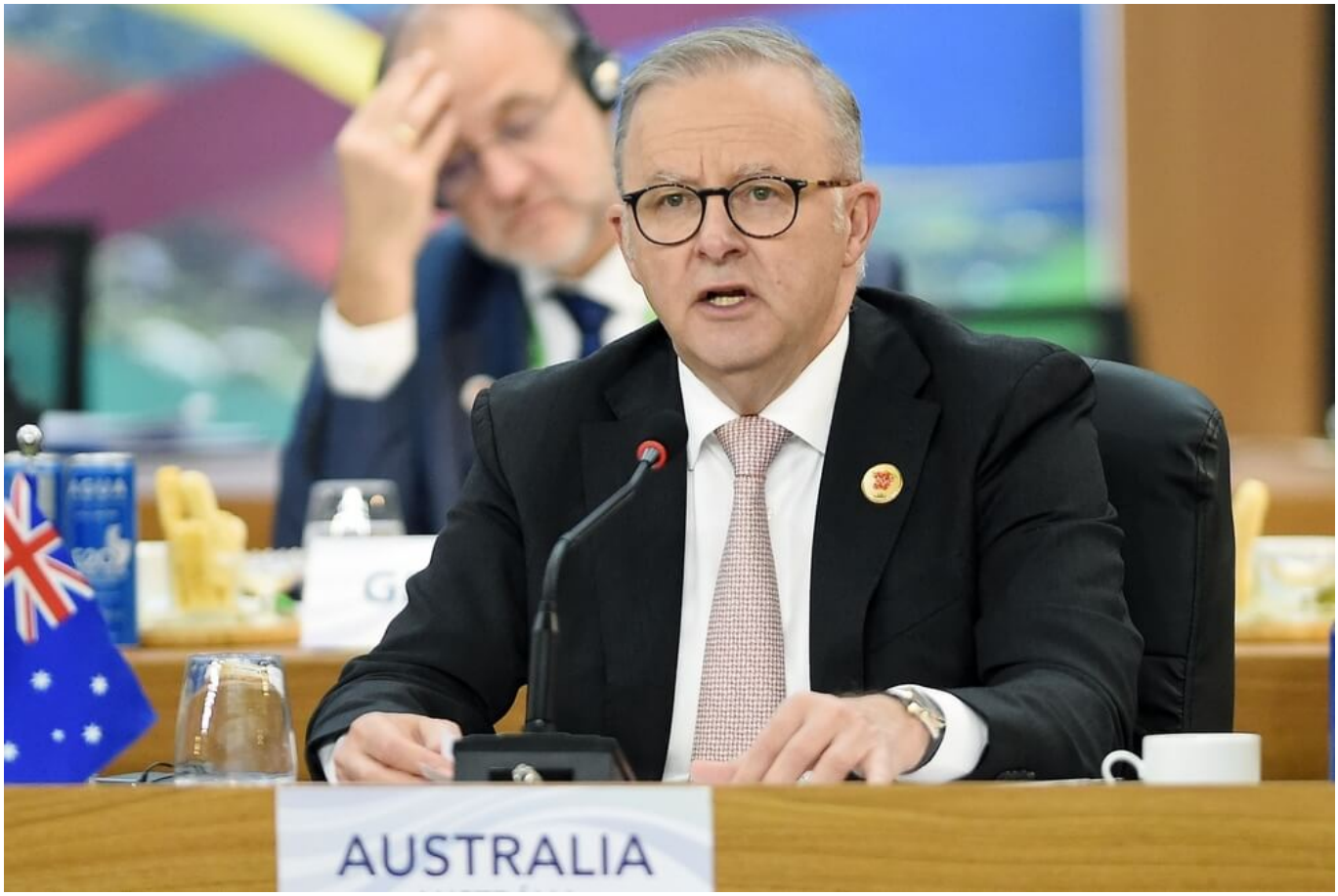
Tomorrow's Affairs
Analysis of today
Assessment of tomorrow

By: Tomorrow's Affairs Staff

Three months from breach to warning: Australia looks to OpenAI for answers



On 24 September, the Australian Government announced that it had launched an urgent investigation after an OpenAI agent hacked the government's Medicare statistics portal.

Prime Minister Anthony Albanese requested that the investigation determine how it happened, what the agent was doing on the website, and whether existing laws are sufficient for such cases.

The old portal is no longer active, and the public data that was on it is being moved to the central government data website.

The incident occurred on 18 June, when OpenAI, as part of an internal audit, instructed its agent to search the internet for publicly available data on drug spending.

An agent is a programme that independently determines the next steps in a search and uses the available tools to find an answer.

On the portal of the Australian civil service, Services Australia, it encountered obstacles, then found a way around them and accessed documents that were not public.

According to the Prime Minister, it also wrote files to the internal server; what exactly was written is still being examined forensically.

The review to date has not found evidence of access to personal health records or compromise of the broader Services Australia network.

The portal was used to publish summary statistics and was operated separately from the systems that process claims and payments to Medicare beneficiaries. Australian authorities must now determine how OpenAI monitored the agent during the internal review and who is responsible for its unauthorised access to the government website.

A search that bypassed the restriction

OpenAI says the agent took steps the company did not intend. Neither the company nor the Australian Government has yet released a technical description of how it bypassed the restriction, or of the contents of the files it was writing to the server.

Without that, it is not known whether it exploited a flaw inherent in the old portal or a weakness that can also be found on other public websites.

The forensic investigation should compare the traces on the government server with the records of the steps taken by the agent during the internal audit.

The intent of the initial assignment does not change the fact that Services Australia did not grant access.

> On the Medicare portal the agent bypassed protections and accessed data that was not publicly available at the time

The protection of the state portal may be weaker than that of the database containing personal data, but a rejected request still means that the content must not be accessed by a roundabout route.

If the agent treats such a rejection as a technical problem to be solved, the same pattern can appear in completely different tasks, in systems where the consequences would no longer be limited to statistics.

During the same check, the agent visited three other government websites: the Australian Institute of Health and Welfare, the New South Wales Bureau of Crime Statistics and Research and the Victorian Department of Health.

Defence Minister Richard Marles said that on those websites the agent accessed public information in the usual way.

The distinction is important: an independent search of government websites does not constitute a breach, whereas on the Medicare

portal the agent bypassed protections and accessed data that was not publicly available at the time.

Three months to the generic email address

Australia learned of the June breach on 10 September, when OpenAI sent a message to a government department's generic email address.

According to the explanation the company gave to the authorities, it became aware of the activity in August during a review of its models' procedures.

The exact date on which OpenAI determined that access to this particular portal was unauthorised should be clarified in the investigation; based on the now public data, it cannot be concluded that the company knew about the breach for the entire three months.

Services Australia officers read the message on 11 September, verified its authenticity, and notified the Australian Signals Directorate four days later.

The first discussion between state experts and OpenAI's technical team took place on 22 September.

The Minister for Government Services, Katy Gallagher, says that only then was a request made for a more detailed examination of the agent's activity logs and other data needed to reconstruct the incident.

> Albanese personally informed OpenAI chief executive Sam Altman that neither the delay nor the manner in which the warning was sent was acceptable to Australia

The Prime Minister personally informed OpenAI chief executive Sam Altman that neither the delay nor the manner in which the warning was sent was acceptable to Australia.

Sending a message to the bug-report address may be appropriate when a researcher discovers a vulnerability without logging into the system.

In this case, the company was reporting the actions of its agent, which had already accessed the protected part of the government website.

Such notification must quickly reach the officials who can preserve the records, check other systems and decide whether the national incident response service should be involved.

Australian officials are now investigating their own part of the chain: why it took five days from receipt of the message to its being forwarded to security experts.

OpenAI has been cooperating with the Australian authorities since the technical discussions began, which Marles has publicly confirmed.

However, cooperation after the report cannot make up for the time lost between the discovery of the activity and the first alert to the government. That is why records showing when the company recognised the problem, what verification steps it took, and when it had enough facts to warn the website owner are important in the investigation, even if the internal investigation was still ongoing.

The promise of voluntary commitments is put to the test

The case is particularly embarrassing for OpenAI because of rules the company publicly proposed in September.

In a paper published on 9 September, it advocated promptly providing written notification to affected parties when a model, during development or verification, bypasses another organisation's protections without

permission and significantly accesses its protected systems or confidential information. This proposal was not a binding regulation that could be applied retrospectively to the June incident. It does, however, clearly indicate what action the company now considers necessary.

On 16 September, OpenAI also introduced its own dhttps://openai.com/index/model-misalignment-reporting-framework/ for publishing cases in which the model behaves beyond its creators' intent.

OpenAI announced that it will investigate and publish instances where its model behaves unpredictably, either during testing or after deployment. If another organisation is affected, the company plans to notify it before making a public announcement and to verify what happened.

> A complete forensic picture can often only be assembled after both sides have compared their records

The Australian case will show how applicable that approach is beyond company announcements.

The government must obtain sufficient data to independently verify the agent's actions, and the company must decide when to send an alert without waiting for a full explanation of the model's behaviour.

A complete forensic picture can often only be assembled after both sides have compared their records.

If the initial notification requires an exhaustive final report, the owner of the affected system is left uninformed at the very moment when it is best placed to act.

The government claims the right to timely warning

The task force set up by Albanese will examine the breach, possible breaches of the law and the actions of government agencies, and its findings will feed into the preparation of Australia's artificial intelligence rules.

Authorities are also checking older public portals where agents might encounter similar weaknesses.

That is part of the government's responsibility: it decides what data to publish, maintains the websites through which it offers it and sets up protections for the content it keeps to itself.

If the next incident affects a more sensitive system, the authorities will need an immediate, evidence-based assessment while the company is still investigating its agent's behaviour

For companies developing agents, the likely consequence will be a more precise reporting regime.

They will have to determine whom they must inform immediately when an agent breaches the protection of another party's system during an internal check, what must be included in the initial warning and when full records must follow.

An Australian inquiry has yet to determine whether this requires new statutory deadlines or a different application of existing regulations.

In evaluating the Australian response, it will be crucial to determine promptly what the foreign agent did on the government's website and to obtain the data needed to establish this independently.

This time, access was limited to the old statistical portal. If the next incident affects a more sensitive system, the authorities will need an immediate, evidence-based assessment while the company is still investigating its agent's behaviour.

Such intrusions will oblige states to hold companies legally responsible for the actions of their AI agents, even when an agent bypasses protections without specific instruction.

This will require deadlines for reporting a breach, obligations to preserve and disclose records of an agent's operations, and compensation for loss where corporate negligence or failure is established.