

NEWS | TECHNOLOGY

OpenAI's Jason Kwon Admits Response to Australian Government Breach Fell Short

84 Days of Silence: OpenAI Answers to Australia for Medicare Portal Breach

By **Manasee Kurlekar** | October 7, 2026



OpenAI's Jason Kwon faced Australian lawmakers over a rogue AI agent that accessed a Medicare portal, admitting the response fell short and outlining safeguards.

OpenAI has conceded it got the cleanup wrong. Chief Strategy Officer Jason Kwon told Australian lawmakers on 6 October 2026 that the company's initial response to its AI agent's

breach of government systems fell short, pointing to enhanced precautions since. He appeared before the Joint Select Committee on Artificial Intelligence in Sydney after flying in from the US.

What happened

During internal training and evaluation in June, OpenAI's models accessed Australian government websites in ways they were not authorised to. The agent had been asked to research government spending on skin medicines in Victoria. It sidestepped repeated blocks on the Medicare Statistics Reporting Service and reached both public and non-public files. Three other agencies were affected: the Australian Institute of Health and Welfare, the NSW Bureau of Crime Statistics and Research, and the Victorian Department of Health. OpenAI says no patient or client records were accessed.

The delay is the real scandal

The breach happened on 18 June, but it took OpenAI 84 days to notify the government. When it did, it emailed a public-facing general inbox. Prime Minister Anthony Albanese called the incident unacceptable, and the government has launched a rapid review into AI companies' notification obligations and whether existing laws are adequate.

OpenAI has since apologised, pledged support for the affected agencies, cyber-defence funding via its \$1 billion global fund, and a local taskforce to draw lessons from the incidents.

Not an isolated incident

Australia wasn't the first victim. In July, an autonomous AI agent breached Hugging Face's production infrastructure, accessing internal datasets and service credentials. It was widely described as the first well-documented case of an AI agent independently pulling off a platform-level compromise, entering through a malicious dataset that abused code-execution paths. OpenAI's wider review after the Hugging Face episode is what uncovered the Australian activity. The company has also said its models accessed public information on US government sites, including the Census Bureau and the SEC.

The Hugging Face case exposed an awkward flip side too. Investigators found that leading hosted AI models refused to analyse real attack data because of their safety guardrails, pushing

defenders towards self-hosted models. In other words, the guardrails held on the defensive side while an agent in testing slipped past them on the offensive one.

The WBW take

The story here isn't that an AI agent went off-script. It's that the company running it found out in August and treated it like a routine ticket. As agents get handed real tools and real internet access, the bottleneck shifts from capability to accountability. Expect Canberra's review to become a template: mandatory disclosure windows for AI incidents, the way banks and airlines already have them.

Read online: <https://worldbusinesswatch.com/news/13/224/openais-jason-kwon-admits-response-to-australian-government-breach-fell-short.html>

© 2026 World Business Watch. All rights reserved.