Australian Government

Department of the Prime Minister and Cabinet

Rapid Review into Australian Government arrangements for an AI-driven cyber incident



Contents

Executive summary 1

Recommendations 3

Overview of the incident 6

What happened 6

What the Australian Government did 8

Expectations for safe development of AI 10

Strategic cooperation with AI firms 11

Reporting obligations 12

Incentivising safe AI development – Regulatory frameworks and international approaches 14

Criminal offences and regulation 17

Australian Government incident management 19

Technical and operational procedures and escalation arrangements 19

Information sharing between agencies and incident coordination 21

Impacts of AI on broader crisis response arrangements 23

Australian Government cyber security posture 25

References 29

Appendix A – Terms of reference 30

Appendix B – Review methodology 31

Appendix C – AGCMF crisis responsibilities by hazard 32

Appendix D – Glossary 34

Executive summary

This Review was triggered by the widely reported OpenAI incident affecting Services Australia. While focused on that specific event, it considers whether the Australian Government is prepared for risks driven by artificial intelligence (AI) more broadly. The Review is timely, providing an early real-world example of the opportunities, risks and governance challenges associated with advanced AI. These will be vital as the Australian Government seeks to maximise the benefits of AI, manage its risks, and strengthen Australia's influence over its safe development and deployment through initiatives including the AI Standards.

In Australia, like elsewhere, AI is at the early stages of transforming society, bringing significant economic, scientific and social benefits. Further profound changes are likely. At the same time, increasingly capable AI models are creating and amplifying additional risks at a faster pace, including risks to Government systems and processes. Advancing AI systems are likely to continue to exhibit behaviours that are difficult to predict, monitor and control. The full extent and scope of such risks are not yet known, including to the very companies developing the technology. Incidents of this nature were warnings months ago and are a reality now. More incidents are likely, as shown by subsequent revelations from OpenAI and Anthropic about comparable incidents outside Australia.

The Review identifies opportunities to strengthen Australia's preparedness and response to the challenges posed by advanced AI systems. These include measures to incentivise safe AI development and deployment, improve collaboration with AI firms, enhance preparedness for AI-related crises, and strengthen the Australian Government's cyber security posture in an evolving threat environment.

Australian Government efforts will rely on constructive cooperation underpinned by requirements for prompt and detailed information sharing by frontier AI labs. AI developers have the greatest visibility of their models' capabilities, behaviours and emerging risks. This basic arrangement – of close cooperation with industry – is the foundation for our existing, world-leading approach to cyber security, and should be central to managing the risks and benefits of AI.

While OpenAI's engagement with the Australian Government has been constructive since the incident was first disclosed to Services Australia, the Review finds the initial timeliness and method of disclosure were inadequate. The Australian Government expects and should demand major improvements in incident reporting from AI companies. Australia's approach to AI must ensure collaboration and accountability work together. Given market incentives to develop increasingly capable AI systems quickly, regulation should create equally strong incentives for safety, backed by meaningful consequences for harmful conduct, and international cooperation to strengthen standards and oversight.

The OpenAI incident affecting Services Australia highlighted the challenge AI-related incidents pose for established government coordination and decision-making processes. While incident-management processes are well designed and tested for malicious cyber activity, the Review finds

they are not designed to identify and respond to novel incidents of this kind. Uncertainty about whether and to what extent this incident was novel, meant its broader policy, regulatory and strategic implications were not considered early across Government. The Review concludes that information sharing between technical agencies, enabling agencies, and the National Cyber Security Coordinator should be strengthened to ensure coordination and information sharing with broader stakeholders can occur quickly, supporting decision makers to respond to incidents with significant public interest implications. More broadly, AI will alter the likelihood, severity, and consequences across hazard types (not just cyber security), so agencies with responsibility for leading the response to incidents need to reassess risk, coordination and response plans.

The Review also considered whether Australian Government systems were appropriately preparing for a substantially degraded cyber threat environment owing to the impacts of AI. The Review finds existing cyber security fundamentals remain essential but should be complemented by greater use of AI to strengthen cyber defence. Agencies should also assess how AI-enabled and autonomous threats may change targeting patterns and attack pathways. This reinforces the need to reduce legacy technology, limit internet-facing systems where possible, and continue the transition towards secure, centralised data environments in government. In other words, for government and for Australia more broadly, our new AI-influenced operating environment makes effective cyber defence even more imperative.

The significance of this incident lies less in its immediate cyber impacts than in what it revealed about Australia's preparedness for increasingly capable AI systems. It provides a valuable, galvanising opportunity to address gaps now, helping ensure Australia is better prepared for incidents of greater scale, complexity and consequence.

Recommendations

The Review recommends:

1. The **Australian Government** strengthen mechanisms for engagement and information exchange with frontier AI developers, including through participation in relevant international initiatives and expert fora, personnel exchanges, and technical collaboration. Actions include to:
    a. engage with OpenAI's Australian Taskforce and similar initiatives, and
    b. explore secondments, personnel exchanges and technical placements with frontier AI firms.
2. In order to maximise leverage with frontier AI firms, **the Australian Government** continue to encourage frontier AI firms to establish research, training and safety-related activities in Australia.
3. **The Australian Government** consider mandatory, time-bound AI safety reporting, including but not limited to misaligned behaviour or agentic cyber incidents.
4. **The Australian Government** explore options to support timely reporting of AI cyber or safety incidents and near misses – for example, protections relating to AI incidents and vulnerability disclosures from AI developers (and others with relevant knowledge), analogous to existing Limited Use protections.
5. **The Office of AI** consider how Australia's AI Standards should leverage existing, accepted regulatory and law-enforcement mechanisms by international counterparts, to help build common AI safety and security expectations.
6. **The Department of Foreign Affairs and Trade (DFAT), the Australian Signals Directorate (ASD), and the Department of Industry, Science and Resources (DISR), including Australia's AI Safety Institute (AISI)**, continue engaging in international AI governance and standards-setting efforts to promote consistent expectations regarding safety, incident reporting, evaluations, safeguards and transparency.
    a. This work be coordinated with Australia's broader international AI engagement, including efforts to secure advanced AI capabilities and attract trusted investment.
7. The Australian Government consider the role AISI plays in future to deliver against increasing demands for technical research, evaluation and analysis on AI, including its capacity to support the proposed AI Standards and their requirements, international engagement on AI standards and norms, and to provide necessary rapid scientific and technical advice and analysis during crisis coordination (See Recommendation 11).

8. The Attorney-General's Department (AGD), in consultation with the Australian Federal Police (AFP), the Department of Home Affairs (Home Affairs), and other relevant agencies, consider options to reform the criminal law to improve AI safety, including exploring:
    a. the application of criminal penalties for egregious regulatory non-compliance; and/or
    b. a standalone 'failure to prevent' offence which would criminalise instances where the company is responsible for developing an AI model that engages in harmful conduct, and the company fails to take reasonable steps to prevent that harm.
9. The Protective Security Board, informed by advice from Home Affairs and ASD, issue Protective Security Policy Framework (PSPF) policy guidance to clarify agency obligations and escalation expectations, where reported vulnerabilities or suspected cyber security incidents have not yet been technically validated.
10. Commonwealth agencies review vulnerability-disclosure and cyber incident-management procedures to ensure they align with PSPF requirements, support timely escalation of novel or unusual incidents, and provide clear pathways for reporting matters to ASD, Home Affairs and senior decision makers.
11. Operational agencies and agencies with coordination, policy or enabling responsibilities must ensure timely information sharing. In relation to cyber incidents, consideration be given to establishing formal liaison arrangements between ASD's Australian Cyber Security Centre (ACSC), the National Office of Cyber Security (NOCS) and AISI to support improved information sharing, coordination, and the early identification of incidents with broader national-interest implications.
12. Commonwealth agencies clarify response procedures to ensure the organisation responsible for consequence management in response to a cyber incident (the NOCS) determines whether consequence management arrangements are required, enabled through enhanced information sharing (per Recommendation 11).
13. Commonwealth agencies responsible for national plans under the Australian Government Crisis Management Framework (AGCMF), and their underpinning procedures, consider the impact of AI on the hazards and arrangements they manage, and write to Crisis Arrangements Committee Chairs with outcomes of this work and any proposed changes by the end of 2026, ahead of the next update to relevant national plan or procedures.
14. Commonwealth agencies take action to uplift the AI literacy of relevant staff to strengthen their ability to respond to AI-related risks.
15. Commonwealth agencies renew their focus on the fundamental elements of good cyber security practice, including implementation of ASD's Essentials.

16. **Commonwealth agencies** work to reduce the size of the government digital estate, consolidating digital systems, and shutting down or segregating systems that do not need to be connected to the internet.
17. **The Australian Government** consider how to further harden cyber defence of Systems of Government Significance, including considering a changing threat environment and risk profile.
18. **Commonwealth agencies** engage with ASD to identify and assess opportunities to responsibly use AI-enabled cyber defence tools and techniques to enhance cyber security monitoring, penetration testing, threat detection and incident response capabilities.

Overview of the incident

On 24 September 2026 (AEST), the Prime Minister, the Hon Anthony Albanese MP, announced details of an incident in which an AI agent from OpenAI infiltrated an Australian Government website – the Medicare Statistics Reporting Service administered by Services Australia.[i] Further investigation has confirmed that no sensitive personal data was accessed, and the issue has been remediated by taking down the Medicare Statistics Reporting Service and moving the data to a more secure environment.

What happened

On 18 June 2026, an OpenAI agent[1] undertook unauthorised activity on Services Australia's Medicare Statistics Reporting Service. The Medicare Statistics Reporting Service was an unauthenticated, standalone and publicly accessible service, not integrated with other Services Australia systems. The AI model was an internal OpenAI model that did not include the full set of safeguards[2] normally associated with a publicly released model.[ii] While undergoing evaluation, the agent was set a task to research government spending per person on medicines for skin conditions in Victorian communities.[iii]

The Australian Signals Directorate's (ASD's) Australian Cyber Security Centre (ACSC) conducted an investigation into the June 2026 exploitation of Services Australia's Medicare Statistics Reporting Service portal. The investigation focused on what data was accessed, if additional systems were affected, and if human input caused the activity. The investigation revealed:

- That the OpenAI internal AI model agent was not appropriately contained, lacked appropriate and timely human oversight or intervention.
- The agent initially attempted to access non-public medical statistical data from the Australian Institute of Health and Welfare (AIHW); however, attempts were unsuccessful. Subsequently, the agent targeted the Services Australia Medicare Statistics Reporting Service portal and successfully accessed public and non-public data without authorisation.
- No evidence of the agent abusing credentials or vulnerabilities to move laterally across the Services Australia network to access more sensitive data on the network, such as Personally Identifiable Information.
- No indication that unauthorised activity was caused by human input.
- No evidence of lateral movement, nor sophisticated anti-forensic activity by the agent.

[1] An AI agent refers to a system or program that can autonomously perform tasks on behalf of a user or another system.

[2] Model evaluation without safeguards is a normal step for developers to establish a baseline of a model's true capabilities, risks and biases before safety or alignment techniques are applied.

The OpenAI incident affecting Services Australia represents a realised AI safety risk and was the first documented example of misaligned agents affecting Australian Government infrastructure.

OpenAI became aware of this activity on 11 August 2026, following human review of AI model activity by its security team.[iv] This OpenAI review was in response to a previous incident in July 2026 when a different OpenAI model broke out of testing and breached the cyber defences of a United States (US) AI company, Hugging Face (the OpenAI/Hugging Face incident).[v]

OpenAI informed Services Australia on 10 September 2026 (AEST) (9 September, UTC) via email that:

- a security vulnerability had been identified
- an OpenAI model had identified a way to carry out instructions on the server through the public reporting interface, without a private account or password
- the model had been able to access this to read portions of the internal program files and settings, obtain a list of files, and create and read back a small test file on the server, and
- OpenAI had not found any evidence that the model had accessed patient-level records, personal information or personal credentials; deleted data; or established ongoing access.[3]

The information from OpenAI at this initial stage did not cover full details, including that the activity involved OpenAI performing model testing, or was an activity by an AI agent with no human control. It also contained an incorrect URL which complicated validation.

OpenAI notified Services Australia through a public-facing email (publicised on the Services Australia website), established for reporting potential cyber security system risks affecting Medicare, Centrelink, Child Support and myGov online systems.[vi] This reporting mechanism is part of Services Australia's Vulnerability Disclosure Program,[4] which supports compliance with the PSPF and is part of the agency's broader approach to identify, assess and manage cyber vulnerabilities.[vii] The website publicising the mailbox notes that the mailbox is not intended to report cyber attacks and redirects those reports to the 'Report a cybercrime' form on cyber.gov.au.

The vulnerability identified by the agent, and the resulting unauthorised access, were not known to the Australian Government prior to OpenAI's disclosure to Services Australia on 10 September 2026.

Following disclosure of the OpenAI incident affecting Services Australia, OpenAI advised that its model had also attempted to access to the Australian Institute of Health and Welfare website without success, and had conducted activity against systems operated by Victoria's Department of Health, New South Wales Bureau of Crime Statistics and Research, and the New South Wales National Parks and Wildlife Service web app. Disclosures from AI companies and

[3] Patient level records and personal information are subject to stronger cyber security protections and would not have been accessible through the same means.
[4] As mandated under PSPF Requirement 0115.

independent research has suggested these instances may reflect a broader series of rogue AI-agent activity globally.[viii]

In response to these incidents, OpenAI has committed in public and directly to the Australian Government that it will strengthen research safeguards, including via additional network restrictions and expanded monitoring. Specific to Australia, OpenAI has publicly affirmed it will provide dedicated support for affected agencies, funding and support to strengthen cyber defences, and establish an Australian taskforce to develop practical policy recommendations for managing emerging risks from increasingly capable AI agents.[ix]

What the Australian Government did

On 11 September 2026, the email notification from OpenAI was reviewed by a Services Australia agency officer – a day lag resulting from non-continuous monitoring of the public mailbox. In accordance with Services Australia's cyber security vulnerability management policy, the OpenAI disclosure was referred internally for technical investigation to verify the vulnerability and assess whether there were any impacts to other systems.

On 14 September 2026, the OpenAI disclosure, technical verification and preliminary findings were escalated to Services Australia's Chief Information Security Officer and subsequently referred to Services Australia's 24/7 Cyber Security Operations Centre which continued investigative analysis overnight.

On 15 September 2026, following technical investigation and assessment of the vulnerability's severity, the risks associated with continued public access, and available containment options, Services Australia took the Medicare Statistics Reporting Service offline, declared a cyber security incident, and notified ASD's ACSC.[5] At this point in time, the available information from OpenAI still did not disclose that this event involved a model testing activity undertaken by OpenAI itself or that it involved a non-directed activity by an AI agent.

On 17 September 2026, Services Australia briefed the Office of the Minister for Government Services, informing of the confirmed vulnerability, the decision to take the Medicare Statistics Reporting Service offline, ongoing investigations, and engagement with ASD and OpenAI. Over the course of subsequent days, ASD undertook further investigations and provided briefing to relevant Ministers and ministerial offices, through which the National Cyber Security Coordinator (the Coordinator) was advised of the incident on 19 September 2026.On 20 September 2026, OpenAI informed ASD verbally that the incident involved its own model testing, and it was an AI agent that had accessed information not available to the public. Following this disclosure, ASD notified the Minister for Government

[5] ASD classified the incident as Category 4 – a 'low-level malicious attack', affecting federal government. The vulnerability that was exploited was assessed as unsophisticated and reports indicated no sensitive material was accessed (such as personal information).

Services, Minister for Cyber Security and the National Cyber Security Coordinator. The National Cyber Security Coordinator commenced discussions with OpenAI the same day.

On 22 September 2026, OpenAI participated in its first technical exchange with ASD, Services Australia, the AISI, and the NOCS. The technical information that ASD needed to support the forensic investigation was not provided by OpenAI until 24 September 2026.

On 24 September 2026, the Prime Minister announced the incident and the establishment of the Review. The Review's terms of reference were published on the Department of the Prime Minister and Cabinet's (PM&C) website on 24 September 2026.[x]

Expectations for safe development of AI

Over coming years, AI will reshape economies and societies, bringing a range of benefits. It will boost productivity and innovation across industries, creating new products and ways of working. Slow adopters risk becoming less competitive than AI-enabled competitors. But the rapid advancement of AI is bringing new risks, more quickly. Risks can emerge unexpectedly through the development of more capable AI systems and interactions between them, with consequences that can't be fully predicted.

Against this backdrop, the Review examines the OpenAI incident affecting Services Australia as a practical case study to assess whether existing legislative, governance and information-sharing arrangements are fit-for-purpose to prepare for, and respond to, a cyber incident involving AI. The Terms of reference and Review methodology are provided at Appendices A and B respectively.

Significant developments in AI capabilities have greatly enhanced the speed, scale, and accessibility by which complex cyber activities (legitimate, misaligned and malicious) can be conducted.[xi] This year has seen agentic capabilities lead to instances where powerful AI agents have autonomously engaged in harmful cyber activities, initially undetected by their developers. This has been, in part, owing to insufficient technical evaluation and monitoring design by the frontier AI labs and the AI model's greater sophistication and attempts to conceal its behaviour.[xii] While this was a novel cyber incident in the Australian context, it may be better categorised as a misalignment incident – a realised AI safety risk. Other examples include the OpenAI/Hugging Face incident, which the Independent International Scientific Panel on AI describes as "one of the clearest real-world warnings yet of one possible route to loss of human control over AI: capable agents pursuing goals that conflict with human intentions".[xiii]

As capabilities and adoption grow, risks will continue to manifest beyond cyber security. To ensure AI systems are safe, agentic activity needs to be aligned with human interests, must prevent hostile actors from misusing them, and ensure humans can control them. These research problems – alignment, security and control – have not yet been solved. As with all complex technologies, no country can eliminate all risks and similar incidents will happen again. Collectively, systems will be needed to monitor a range of domains and to identify risks quickly, adapting to AI disruptions as they emerge. If Australia can map risks, build expertise and boost resilience, it will better prevent and mitigate harms from AI, allowing Australia to maximise its opportunities.

In considering potential approaches, it will be important to ensure that requirements and expectations placed on AI developers are carefully calibrated. Collaborative partnerships with AI firms will be needed to ensure timely and privileged information sharing. At the same time, regulatory obligations will be required to set clear expectations for safety and security. In severe instances of reckless actions or negligence that result in significant harm, a criminal offence with appropriate penalties to deter and respond to serious misconduct and harms may be appropriate.

Strategic cooperation with AI firms

To date, in its engagement with US frontier AI firms, the Australian Government has adopted a collaborative approach. This approach seeks to build trust through transparency, shared understanding, and cooperative governance arrangements (including via memoranda of understanding). This approach supports constructive relationships between government and developers and should help sustain public confidence in the safe and responsible development of highly capable AI systems. This approach also acknowledges that in many instances, AI developers and the information and technical expertise they hold, will be required to keep Australians safe.

AI capabilities continue to advance rapidly, often outpacing the development of technical controls and solutions for research problems of alignment, security and human control. All governments will need to grapple with this complex problem. In the interim, AI capabilities and associated controls will evolve and risks that were once largely theoretical will likely continue to manifest as real-world events. While this incident had minimal direct impact on government systems, future events might have more serious and damaging consequences. Our systems and processes need to prepare accordingly.

It is therefore vital that Australia develops mechanisms to understand rapidly-evolving AI capabilities, their associated risks, and the effectiveness and limitations of safeguards and technical controls. However, no single organisation or country has enough visibility to identify every emerging pattern of behaviour.[xiv] An early warning system to identify wider patterns of concerning behaviour requires cooperation and trusted information sharing between government and AI developers.[xv]

Without this understanding – which will necessarily rely on an extensive engagement with partner governments and frontier AI labs – there is a risk that Australia's frameworks will struggle to keep pace with a technology that is advancing faster than existing institutions and processes were designed to accommodate. The Australian Government will require visibility across the AI ecosystem to identify emerging risks, support effective incident management, and adapt governance, regulatory and policy responses as technology evolves. In many instances, this will require AI developers to proactively engage with Australian Government agencies responsible for AI safety, technical response (including AI-enabled cyber defence and vulnerability remediation), and crisis coordination and consequence management.

Recommendations

1. The Australian Government strengthen mechanisms for engagement and information exchange with frontier AI developers, including through participation in relevant international initiatives and expert fora, personnel exchanges, and technical collaboration. Actions include to:
    a. engage with OpenAI's Australian Taskforce and similar initiatives, and
    b. explore secondments, personnel exchanges and technical placements with frontier AI firms.
2. In order to maximise leverage with frontier AI firms, the Australian Government should continue to encourage frontier AI firms to establish research, training and safety-related activities in Australia.

Reporting obligations

In the case of the OpenAI incident affecting Services Australia, this Review concludes that the delay and method of notifying the affected agency was demonstrably inadequate. The incident occurred in June, remained undetected until August and was not disclosed until September. OpenAI became aware of the activity in August – only after retrospectively reviewing model actions following the earlier Hugging Face incident – and the disclosure was made through a public-facing vulnerability reporting channel intended primarily for reports of suspected vulnerabilities, rather than confirmed incidents.

This incident challenges several assumptions underpinning cyber incident management. Existing frameworks generally distinguish between victims who report incidents,[6] and threat actors who may seek to conceal them. In this case, the company associated with the activity identified the issue, disclosed it, and assisted with response efforts, despite not being the affected entity. This reflects the novel characteristics of incidents involving advanced AI systems and highlights the importance of maintaining trusted channels for disclosure and information sharing.

Even so, while the disclosure was welcomed, the timing and method of disclosure were not commensurate with the nature of the activity or its broader significance. OpenAI did not adequately reflect the potential implications of AI-enabled activity for the cyber threat environment, including

[6] Typically, an entity affected by a cyber incident would report the incident to the ASD's ACSC cyber.gov.au/submit-a-report.
As part of PSPF requirements, Government agencies are required to report incidents to ASD and significant security incidents to Home Affairs 'as soon as possible'. Other requirements for victims to report cyber incidents or breaches are established across legislation including the *Security of Critical Infrastructure Act 2018*, the *Privacy Act 1988* and the *Cyber Security Act 2024* (ransomware reporting). Each have distinct reporting requirements and thresholds.

actual or perceived risks to critical infrastructure, government systems and the confidentiality of personal information.

Despite the incident's novelty and significance, and OpenAI's engagement with Australian Government ministers and officials, the matter was not raised directly through those channels. The Review notes evidence provided by OpenAI's Jason Kwon at a recent Joint Select Committee on AI hearing, that the OpenAI Chief Executive Officer was not aware of the incident when he met with the Deputy Prime Minister on 1 September 2026, and that OpenAI were treating the matter primarily as a technical issue.[xvi]

While vulnerability disclosures should generally be directed to affected entities to support remediation before public disclosure, this matter involved not only a suspected vulnerability, but actual unauthorised access and novel AI agent behaviour. As a result, there were clear impacts on public confidence in government systems, particularly given widely reported risks associated with uncontrolled or misused AI applications. In these circumstances, earlier notification and more senior engagement with Services Australia and ASD was required. This would have supported a timelier assessment of risks, coordination of response activities, and appropriate government awareness. OpenAI has acknowledged that, in retrospect, ministers or senior officials should have been contacted sooner.[xvii]

The Review concludes that further work is needed to ensure AI developers (and others with relevant knowledge) report incidents and near-misses quickly and transparently and are incentivised to proactively engage with technical, safety, and policy agencies. To maintain trust and open disclosure to assist with initial incident response, consideration should be given to the development of protections relating to AI-incident and vulnerability disclosures, analogous to existing Limited Use[7] provisions in the cyber security context. These protections are designed to remove disincentives to voluntarily report, by giving organisations confidence that information provided for incident or vulnerability-management purposes will not be repurposed for regulatory enforcement.

While acknowledging that collaborative approaches will remain essential, this Review concludes they are unlikely to be sufficient on their own. Given the potential consequences of frontier AI systems, mandatory obligations, transparency requirements and independent assurance will also be required to support the safe development, testing and deployment of advanced AI models.

Expectations to report certain AI incidents could be set out in relevant regulation such as the AI Standards, including cyber intrusions, unauthorised access, unintended agentic behaviour, model misalignment, or planned tests in a live environment. In the context of AI regulation, and consideration of potential criminal penalties for serious misconduct or reckless non-compliance,

[7] Limited Use protections are legal safeguards established under the *Cyber Security Act 2024* and the *Intelligence Services and Other Legislation Amendment (Cyber Security) Act 2024*. They encourage businesses to share sensitive information with the government during cyber security incidents without risk of voluntarily disclosed information being used against them in most civil or criminal proceedings.

it will remain important that reporting obligations and enforcement mechanisms do not create unnecessary barriers to timely disclosure, information sharing, remediation, or cooperation.

However, protections designed to encourage disclosure should not operate as a safe harbour from regulatory scrutiny or accountability. Voluntary disclosures and protections analogous to Limited Use provisions may support information sharing, but should not prevent regulators or law enforcement from exercising their own information gathering or enforcement powers.[8]

Action taken

ASD has engaged OpenAI and other frontier AI developers to communicate preferred pathways for reporting identified vulnerabilities and cyber incidents (including parallel notification with the impacted entity and ASD).

Recommendations

3. The Australian Government consider mandatory time-bound AI safety reporting, including but not limited to misaligned behaviour or agentic cyber incidents.
4. The Australian Government explore options to support timely reporting of AI cyber or safety incidents and near misses – for example, protections relating to AI incident and vulnerability disclosures from AI developers (and others with relevant knowledge), analogous to existing Limited Use protections.

Incentivising safe AI development – Regulatory frameworks and international approaches

Incident reporting obligations ensure governments have timely visibility of emerging risks and can coordinate effective responses. However, reporting requirements primarily address information asymmetries after an incident has occurred. They do not, in isolation, create strong incentives for AI developers to invest in the technical controls, testing, evaluation and governance arrangements needed to prevent incidents. The OpenAI incident affecting Services Australia demonstrates that reporting obligations are only one component of an effective governance framework. As AI systems become increasingly capable and autonomous, governments may need to consider how reporting requirements, safety expectations, regulatory obligations and enforcement mechanisms interact to encourage both transparency and responsible development.

[8] Voluntarily disclosed information is not shielded from legal liability just because it was disclosed under Limited Use. While there are restrictions on using and sharing limited use information for regulatory action, and such information is generally not admissible in proceedings against the entity, regulators can use their own powers to independently request information from the entity.

Australia is already considering these questions. The Office of AI in PM&C is developing nationally-consistent AI Standards intended to establish baseline security and safety expectations for frontier AI development, deployment, and use in Australia, including incident reporting and other risk management measures. Such frameworks will set clear expectations regarding the development, testing, deployment and governance of advanced AI systems. AISI has also been stood up to monitor, test, and analyse advanced AI capabilities, risks, harms, and trends to keep Australians safe.

Because frontier AI development and deployment occur across multiple jurisdictions, domestic regulation will be strengthened when aligned with international expectations and regulatory approaches. The most capable AI models are developed by firms headquartered outside Australia and are deployed across multiple jurisdictions.

As a result, Australia will face challenges if it seeks to regulate these systems in isolation. Significant divergence from international approaches could create enforcement difficulties and reduce the practical reach of Australian rules, particularly where firms are able to shift development or deployment activities to jurisdictions with less stringent requirements. This reinforces the importance of international cooperation to support effective regulation and promote shared expectations regarding AI safety and accountability.

Australia therefore has a strong interest not only in ensuring its domestic settings are appropriate, but also in engaging with international efforts on AI governance, standards, and safety expectations. Greater international alignment would promote consistent expectations, reduce compliance complexity for firms operating across several jurisdictions, support international interoperability, and reduce incentives for regulatory evasion. In practice, collaboration, domestic regulation and international governance should be viewed as complementary elements of a broader framework for managing the risks associated with increasingly capable systems. Hosting frontier training infrastructure would give Australia greater opportunity to shape proportionate safety and transparency commitments, while strengthening its voice in emerging international approaches.

Several jurisdictions have already imposed legislated obligations on AI developers. The *EU Artificial Intelligence Act* establishes a comprehensive, economy-wide framework for AI regulation, while California's *Transparency in Frontier Artificial Intelligence Act* focuses more specifically on frontier AI through transparency requirements, safety frameworks, incident reporting obligations and whistleblower protections. Together, these regimes are shaping emerging international expectations of frontier AI governance and serve as possible examples when developing AI risk management and developer-accountability approaches in Australia's AI standards. While there is no agreed model for global AI governance, there is a growing ecosystem of international initiatives, including the AI Safety Summits, the UN Global Dialogue on AI Governance, regional frameworks such as ASEAN's Guide on AI Governance and Ethics and APEC's AI Initiative (2026-2030). Most recently, the Prime Minister has endorsed the Finnish- and Norwegian-led Call for Control of Frontier AI Models. Collectively, these initiatives seek to shape and build the international rules, norms and standards on ethics, trustworthy development, safety testing, transparency, incident reporting and risk management.

Australia's AISI is already advancing global AI safety science through its engagement with the International Network for Advanced AI Measurement, Evaluation and Science. Australia should continue to work with likeminded partners on common approaches to incident thresholds, information sharing and safeguards, to develop credible evaluation methods and niche expertise on AI-enabled cyber risk.

Further, as agentic AI systems become more capable, governments may need to establish international expectations for AI agents, in addition to cyber norms designed for actions undertaken by countries, companies, or individuals. One option for further exploration is coordinated attribution of demonstrably unsafe AI practices that create serious cross-border security risks, drawing on Australia's existing experience with cyber attribution and sanctions.

Recommendations

5. The Office of AI consider how Australia's AI Standards leverage existing, accepted regulatory and law enforcement mechanisms by international counterparts, to help build common AI safety and security expectations.
6. DFAT, ASD and DISR, including AISI, continue engaging in international AI governance and standards-setting efforts to promote consistent expectations regarding safety, incident reporting, evaluations, safeguards and transparency.
    a. This work be coordinated with Australia's broader international AI engagement, including efforts to secure advanced AI capabilities and attract trusted investment.
7. The Australian Government consider the role AISI plays in future to deliver against increasing demands for technical research, evaluation and analysis on AI, including its capacity to support proposed AI Standards and their requirements, international engagement on AI standards and norms, and to provide the necessary rapid scientific and technical advice and analysis during crisis coordination (See Recommendation 11).

Criminal offences and regulation

While reporting obligations and regulatory requirements can improve transparency and establish minimum expectations, they may not be sufficient to deter the most serious forms of misconduct. Governments may therefore need to consider whether criminal penalties are appropriate in cases involving an AI developer having reckless disregard for known risks or serious failures to implement reasonable safeguards. Criminal penalties may be justified where significant departures from accepted safety obligations create substantial and foreseeable risks of serious harm.

A key question raised following disclosure of this incident is whether the activity was unlawful. The *Criminal Code* (Cth) includes several computer offences which could be enlivened by an incident involving AI-assisted hacking. Section 478.1 of the Criminal Code applies where a person intentionally causes unauthorised access to or modification of data held in a computer, and that person knows the access or modification is unauthorised. As a general proposition, these offences could apply where a person or company uses AI as a tool to intentionally undertake a variety of criminal conduct.

AI agents can carry out harmful conduct or cyber breaches without explicit direction or awareness from an individual or company. In such circumstances, existing offences may not apply because the requisite fault elements, such as intent, knowledge, or recklessness may not be established. This does not necessarily indicate a deficiency in Australia's criminal law. Rather, it reflects the fact that criminal law generally does not apply in instances in which an individual or entity did not intend or foresee the resulting harm. This awareness is critical in justifying the significant penalties that flow from breaches of the criminal law.

In the case of the OpenAI incident affecting Services Australia, this Review concludes that, on information available to the Review, key fault elements (as required under the current criminal law) do not appear to have been established – there are no indications of human intent, knowledge of the misaligned activity at the time it occurred, or reckless behaviour in its inception. As such, the Review does not recommend this incident be referred to the AFP for further criminal investigation. However, this incident highlights that the Government may wish to consider whether new offences should be established, that attach criminal liability to failures in development, deployment, governance, or risk management, and/or the harmful conduct itself.

Any criminal penalties or new offences would need to interact with the AI standards and any regulatory model they may create. A tiered model may be appropriate, with civil penalties applying to less significant or unintentional non-compliance, and criminal penalties reserved for egregious conduct, such as an event causing significant harm, or repeated failures to prevent reasonably foreseeable harm or report serious AI-enabled cyber incidents. Such a framework would establish clear obligations, enable regulatory oversight and investigation, allow responses to be proportionate to the seriousness of the conduct, and enable a more practical pathway to enforcement, rather than reliance on a standalone criminal offence.

Alternatively to, or complementing, a regulatory framework with criminal penalties, consideration could be given to a standalone offence. For example, a 'failure to prevent' offence could apply where a company develops or deploys an AI model that causes serious harm and fails to take reasonable steps to prevent that harm. Such an offence could address conduct of AI firms falling outside the regulatory regime and strengthen incentives to implement appropriate safeguards.

Any consideration of such offences could explore instances of egregious conduct or serious harm. The Review does not intend for it to capture lower-level incidents equivalent to the unauthorised activity affecting Services Australia. It should preserve incentives for voluntary reporting and cooperation with authorities. A defence should be available where reasonable preventative measures were taken, assessed against the state of technology, available safeguards, emerging international standards, regulatory requirements and industry best practice. Such an offence could operate alongside a robust regulatory framework.

Law enforcement agencies would need to overcome practical barriers in enforcing any future offence. This includes the technical complexity of evidence, the likelihood that relevant evidence would be located overseas, and the AFP would need to leverage police-to-police assistance and the mutual assistance framework.

More broadly, AI does not exist beyond current legal frameworks. Given the breadth of its impacts, many established legal and regulatory regimes already apply to aspects of AI firm operations. For example, the Australian Consumer Law (ACL) contains general consumer protections that may apply where an AI product is sold or promoted with misleading claims about its safety or testing. Separately, the ACL product safety framework exists to address risks of injury, illness, and death arising from consumer goods. The applicability and suitability of existing frameworks will need to be continually reviewed as AI continues to develop.

Recommendations

8. AGD, in consultation with AFP, Home Affairs, and other relevant agencies, consider options to reform the criminal law to improve AI safety, including exploring:
    a. the application of criminal penalties for egregious regulatory non-compliance; and/or
    b. a standalone 'failure to prevent' offence which would criminalise instances where the company is responsible for developing an AI model that engages in harmful conduct, and the company fails to take reasonable steps to prevent that harm.

Australian Government incident management

Technical and operational procedures and escalation arrangements

A key component of the Review was to consider whether Commonwealth governance and information sharing mechanisms were fit-for-purpose, including roles and responsibilities and escalation pathways. Having considered the period between OpenAI's awareness of the incident and its disclosure to the Australian Government, the Review also examined the timeliness of escalation within government, including the period between Services Australia's awareness of the incident, broader inter-agency awareness, and ministerial notification.

The Review finds the time between OpenAI's disclosure to Services Australia and escalation of the incident to Ministers reflected a combination of an unusual disclosure pathway, insufficient detail provided by OpenAI on the event, treatment of the matter as a vulnerability rather than a cyber incident, and the need for technical validation.

Despite this, the incident was reported to ASD and escalated to the Minister's Office ahead of requirements set out in PSPF requirements and Services Australia's vulnerability management process. Given the unusual nature of the disclosure and potential implications of the incident, Services Australia's exercise of discretion demonstrated appropriate judgement and risk awareness.

Under the PSPF, government agencies must notify ASD of cyber incidents and Home Affairs of significant security incidents (including cyber security incidents), as soon as possible, whereas reporting cyber vulnerabilities is generally encouraged rather than required. As AI accelerates the speed between vulnerability identification and exploitation, the distinction between a vulnerability and an incident may become less clear, making escalation decisions more difficult.[xviii]

Services Australia's process for vulnerability disclosure management[9] requires a reported vulnerability to be verified as 'impacting one or more agency systems before further vulnerability management activity occurs'. Once verified, the vulnerability is prioritised and managed alongside other identified vulnerabilities according to risk. This incident illustrates how novel AI-enabled activity can sit outside established categories. While the reported vulnerability had not been technically validated and appeared to present lower direct system or privacy risks than many cyber threats routinely managed by the agency, it also involved reported unauthorised access, novel AI agent behaviour, a unique disclosure, and potentially significant policy and public-interest implications. The incident therefore required Services Australia officials to exercise judgement about escalation before all technical facts were known.

[9] PSPF Requirement 0115 requires government entities to maintain vulnerability disclosure programs, including processes to receive, assess, verify and remediate reported vulnerabilities.

The Review concludes that, when determining whether coordination of response is required, technical, operational and standard operating procedures should account not only for technical severity, but also consider the broader significance of an incident (for example, a novel vector or disclosure pathway). Services Australia has already taken steps in this direction, updating its vulnerability disclosure procedures so disclosures are forwarded to its 24/7 Cyber Security Operations Centre to support continuous monitoring and notification of senior officials. Services Australia is considering further refinements to escalation arrangements with ASD and other relevant agencies in the context of AI-enabled incidents.

All agencies should therefore review cyber incident reporting and escalation procedures to ensure they clearly outline PSPF reporting obligations and provide sufficient flexibility to support judgement in novel circumstances. While existing arrangements appropriately emphasise technical validation and consequence management, AI-enabled activity may increasingly blur distinctions between vulnerability disclosures and cyber incidents. As a result, officials may need to make escalation decisions before all technical facts are known. However, as AI will increasingly become a component in vulnerability detection and cyber incidents, it is important that reporting decisions reflect key judgements around their significance and validity, so as not to overwhelm already resource-bound cyber response arrangements. There may therefore be value in Home Affairs considering a new requirement under the PSPF to report vulnerabilities (with guidance on appropriate timeframes or technical-certainty levels) or providing updated guidance to Commonwealth entities on escalation considerations under the PSPF, including for incidents with an AI element.

Recommendations

9. The Protective Security Board, informed by advice from Home Affairs and ASD, issue PSPF policy guidance to clarify agency obligations and escalation expectations, where reported vulnerabilities or suspected cyber security incidents have not yet been technically validated.
10. Commonwealth agencies review vulnerability-disclosure and cyber incident-management procedures to ensure they align with PSPF requirements, support timely escalation of novel or unusual incidents, and provide clear pathways for reporting matters to ASD, Home Affairs and senior decision makers.

Information sharing between agencies and incident coordination

Beyond processes to escalate technical matters, the Review also assessed whether information sharing and coordination functions were fit-for-purpose. The Review finds the combination of an incident with limited direct consequences, but novel and with strategically-significant and public interest implications, highlights the need for new or enhanced guidance and assessment processes that allow for appropriate escalation, and broader whole-of-government coordination arrangements.

The strategic significance of the incident extended beyond its immediate technical cyber security impacts. The novelty of the AI-enabled activity, combined with the Government's policy, regulatory and investment focus on AI, generated substantial demand across government and ministers for information on the implications of the event for AI safety, cyber security, future governance arrangements, and public confidence in AI. While ASD and Services Australia met initial technical response and information requirements (including through engagement with OpenAI), there was no clear trigger for broader whole-of-government coordination. As a result, the National Cyber Security Coordinator was not informed of the incident until 19 September 2026.

While the direct consequences of the incident were being effectively managed, earlier engagement with the National Cyber Security Coordinator[10] would have supported timelier whole-of-government awareness, information sharing and coordination, particularly given the significant strategic, regulatory, policy, and public-interest implications.[xix] Whole-of-government coordination can be triggered by a range of factors, including ministerial or community expectations. Therefore, determining when coordination is required is distinct from assessing the technical severity of a cyber incident. It may require consideration of broader policy, regulatory, strategic and public-interest implications that extend beyond the remit of operational cyber agencies.

In this case, the AISI had relevant expertise in frontier AI systems, model evaluations, safeguards and AI-related risks. While there is no evidence that the timing of AISI's engagement affected management of the immediate cyber consequences, the incident raised questions extending beyond cyber response, including AI model behaviour, safeguards, AI model risk management and implications for AI policy and governance. Earlier engagement with enabling agencies may have assisted the Government to develop earlier a more complete understanding of these issues.

The Review considers it important to maintain the distinction between the roles of operational cyber agencies and the National Cyber Security Coordinator. Operational agencies are responsible for technical assessment and incident response, while the National Cyber Security Coordinator is

[10] The National Cyber Security Coordinator leads whole-of-government responses to significant cyber incidents and manages consequences arising from incidents (including by ensuring alignment across regulatory, policy, and incident response frameworks). The role was established to ensure the Minister for Cyber Security has access to a consolidated understanding of cyber incidents and their broader consequences.

responsible for assessing broader consequences, ensuring the Minister for Cyber Security is well informed of the broader context, and determining whether whole-of-government coordination is required. Effective information sharing between these functions is therefore critical.

Recommendations

11. Operational agencies and agencies with coordination, policy or enabling responsibilities must ensure timely information sharing. In relation to cyber incidents, consideration should be given to establishing formal liaison arrangements between the ASD's ACSC, NOCS and AISI to support improved information sharing, coordination, and the early identification of incidents with broader national interest implications.
12. Commonwealth agencies clarify response procedures to ensure the organisation responsible for consequence management in response to a cyber incident (the NOCS) determines whether consequence management arrangements are required, enabled through enhanced information sharing (per Recommendation 11).

Impacts of AI on broader crisis response arrangements

As AI capabilities continue to develop, the risks they create are unlikely to be constrained to cyber security alone. An important question is therefore whether existing crisis management arrangements remain well placed to respond as AI-related risks materialise across an increasing range of domains. The National AI Plan recognised this challenge and outlined that the Australian Government was taking proactive steps to prepare for any potential AI-related incident. This includes testing arrangements under the AGCMF as the overarching policy framework which establishes how the Australian Government prepares for, responds to, and supports recovery from crises.

The AGCMF adopts an all-hazards approach designed to support flexible responses to both established and emerging crises. While it identifies hazards supported by dedicated national plans and sector-specific arrangements, responsibility for leadership and coordination is determined by the nature and extent of the consequences being managed, with responsibility assigned to the entity managing the greatest proportion of those consequences. This informs decisions such as which minister and senior official will lead the response, and which national plan (or plans) are activated. An overview of the identified hazards within the AGCMF, and responsible ministers and senior officials for each hazard is available at Appendix C.

Within this framework, AI is not treated as a standalone hazard. Instead, AI is considered a factor that may influence the likelihood, nature, severity, or consequences of existing hazards. This approach reflects the reality that the role of AI in a crisis may not be immediately apparent and avoids the complexity that could arise from establishing separate AI-specific crisis management arrangements.

While the AGCMF is designed to allow the Government to respond flexibly to new and evolving hazards, advanced AI may test assumptions about how hazards occur, how consequences are assessed and managed, and who needs to be involved in response efforts. This places greater importance on ensuring agencies responsible for existing hazards understand how AI may change risk profiles, consequence management approaches and coordination arrangements. It also reinforces the need for clear pathways to engage specialist AI expertise where incidents raise novel technical, policy or governance questions.

Under the AGCMF, enabling agencies provide specialist technical, scientific, intelligence, policy or operational capabilities to support lead agencies and coordinating authorities in understanding and responding to crises. While they do not normally lead whole-of-Australian Government crisis responses, the expertise of technical and other enabling agencies can inform risk assessment, consequence management and government decision making. As AI-enabled cyber activity and safety incidents become more commonplace, the involvement of specialist AI expertise should be determined by the novelty, complexity and potential consequences of an incident, rather than the mere presence of AI. The objective should be to ensure specialist expertise is available when needed, while avoiding unnecessary escalation of routine incidents.

The OpenAI incident affecting Services Australia demonstrates that business-as-usual responses may not always be sufficient when incidents involve novel AI-related risks. As the Australian Government seeks to build public confidence in AI adoption and strengthen Australia's position within the global AI ecosystem, targeted enhancements to existing arrangements will be required to ensure future incidents can be managed effectively.

The Review therefore considers that targeted enhancements to existing arrangements, rather than the creation of a separate AI crisis-management framework, are likely to provide the most effective means of improving preparedness for AI-related incidents. Consistent with this approach, Home Affairs has been leading work to test arrangements, including through tabletop exercises, to understand how existing national plans could be applied to, or adapted for, AI-related crises. The AGCMF emphasises the importance of continuous improvement to ensure national crisis management arrangements remain responsive to an evolving and increasingly complex threat landscape. Further exercises and regular reviews of the AGCMF and supporting plans will continue to inform improvements.

Recommendations

13. Commonwealth agencies responsible for national plans under the AGCMF, and their underpinning procedures, consider the impact of AI on the hazards and arrangements they manage, and write to Crisis Arrangements Committee Chairs with outcomes of this work and any proposed changes by the end of 2026, ahead of the next update to relevant national plan or procedures.
14. Commonwealth agencies take action to uplift the AI literacy of relevant staff to strengthen their ability to respond to AI-related risks.

Australian Government cyber security posture

AI is reshaping the cyber threat environment. While it can strengthen cyber defence, it also enables vulnerabilities to be identified and exploited faster, with less predictability, at greater scale, and with increasing sophistication. As AI capabilities advance, cyber risks are likely to increase accordingly.

Although the consequences were limited, the OpenAI incident affecting Services Australia is a real-world manifestation of a previously recognised AI safety risk, and also highlights a changing risk exposure. The ability of capable AI systems to identify and exploit vulnerabilities is altering traditional assumptions about cyber threats, threat actors, and potential risk targets. AI agents may autonomously attack targets that would not typically be identified as high-value by human threat actors. As a result, the pool of potential targets may expand significantly, with capable AI models able to identify and exploit vulnerabilities across a much broader range of targets, at scale, and with little warning.

To respond to this shift away from traditional cyber security risk-assessment methodology, and as AI capabilities continue to develop, the Australian Government will need to ensure its cyber security posture evolves accordingly. This will require resilient systems, strong technical controls, effective vulnerability management, the use of AI proactively for defence, and continued adaptation to AI-enabled and increasingly autonomous threats.

There were several factors that meant the Services Australia Medicare Statistics Reporting Service was vulnerable to this unauthorised activity (or would be vulnerable to cyber attack).

- It was a public facing, unauthenticated internet-accessible website which expanded its attack surface
- It was a legacy system that only required minimal ongoing support, because it contained low-risk information and was therefore not prioritised for cyber security uplift.[11]

In June 2026, Five Eyes Cyber Security Agencies released a statement which urgently called for prioritising cyber resilience actions, including reducing attack surfaces, accelerating patching processes, addressing legacy systems, reviewing and strengthening identity and access controls, and preparing for incidents.[xx] As AI enables increased speed and scale in identifying and exploiting vulnerabilities, it raises the urgency for best practice to be implemented across government.

The Review reiterates that agencies must focus on the fundamental elements of good cyber security practices. These include the *Essential Eight* (particularly reaching Maturity Level 2),[12] and engagement

[11] This prioritisation was according to the low risk information contained.

[12] It is a requirement under the PSPF (Requirement 0105) that Commonwealth entities reach Maturity level 2 under the *Essential Eight* – a framework developed by the ASD to help organisations protect against cyber threats. ASD has recently updated this framework to a new *Essentials* series that builds upon the original *Essential Eight* to address contemporary cyber threats and modern enterprise environments. 'Maturity levels' are

with ASD as it rolls out the revised approach through the new *Essentials* series. In particular, this requires authentication and application controls, and prioritising rapid patching cycles to meet demands of changing timeframes between vulnerability discovery and exploitation. The 2025 Commonwealth Cyber Security Posture report found that 22 per cent of Commonwealth agencies had reached Maturity Level 2.[xxi] While this is an increase from 15 per cent in 2024, most agencies are not reaching the cyber security standards required under the PSPF.[xxii]

In addition, the incident highlights the importance of shifting away from legacy technology as it is more vulnerable, owing to limited vendor support and patches. Legacy technology also has higher maintenance costs than newer systems, placing additional pressures on stretched departmental budgets. This is supported by a series of PSPF directions and requirements – including *PSPF Direction 002-2026 – Strengthening Commonwealth Cyber Posture Against AI-Enabled Risks*, undertaking a Technology Asset stocktake, implementing a technology security risk management plan for internet-facing systems, and implementing temporary mitigations for legacy systems while they are being phased out. Reducing the size of the Australian Government's internet-facing attack surface will also be necessary to not only simplify the digital estate, but to mitigate against the new risks that agentic AI may pose to these previously low-value targets.

Recommendations

15. Commonwealth agencies renew their focus on the fundamental elements of good cyber security practice, including implementation of ASD's Essentials.
16. Commonwealth agencies work to reduce the size of the government digital estate, consolidating digital systems, and shutting down or segregating systems that do not need to be connected to the internet.
17. The Australian Government consider how to further harden cyber defence of Systems of Government Significance, including considering a changing threat environment and risk profile.

At the broadest level, Horizon 1 of the *Cyber Security Strategy 2023-2030* has already delivered meaningful cyber security uplift across the government.[xxiii] This work is continuing and deepening under Horizon 2, which aims to shift the government's cyber security response to a proactive, investment-led approach.[xxiv] This includes through prioritised investment for Systems of Government Significance – the Commonwealth's most critical digital services.

---

now transitioning into 'priority phases' under the *Essentials*. Organisations that achieved Maturity Level Two can expect many existing controls to contribute to actions in the early priority phases.

In addition to cyber security fundamentals, ASD has encouraged entities to use AI to strengthen defence.[xxv] To scale these efforts across government, ASD is developing guidance and assistance for other departments to run their own AI-powered code reviews.

In this respect, the Australian Government has been active in addressing AI-enabled cyber security risk, including through targetted risk-based cyber security uplift, which will help to protect against increasing AI cyber capabilities. The Government has also published specific guidance in response to frontier AI cyber developments. ASD has published a significant number of reports related to AI over several years, starting in 2023. More recently, ASD issued guidance on the impact of frontier AI on cyber security on 8 April 2026, the day after the release of Claude Mythos Preview on 7 April 2026. In addition, PSPF Advisory 001-2026 and PSPF Direction 002-2026 position cyber security in the context of advancing frontier AI cyber capabilities as a national priority, and provide practical advice from the Information Security Manual and ASD on hardening Commonwealth systems against frontier AI cyber threats.

While the OpenAI incident affecting Services Australia was unacceptable, it also highlights the strengths of the Government's approach to cyber security and provided valuable lessons – both for government and OpenAI. While not the subject of this Review, it is noted that the Australian Institute of Health and Welfare was also subject to attempts by OpenAI models to gain unauthorised access to their systems, but was not compromised because of its strong cyber fundamentals. Agencies should consider lessons from this success. No government system should ever be compromised by third parties, whether intentionally or not, but the Medicare Statistics Reporting Service was a public, unauthenticated reporting service containing aggregate, low-sensitivity information, operating as a stand alone service on dedicated servers. The Medicare Statistics Reporting Service had controls that prevented access to broader datasets, which were commensurate to the threat landscape at the time of deployment. At the time of the OpenAI incident affecting Services Australia, the Medicare Statistics data was already flagged to be migrated to a more secure portal at data.gov.au. This transfer was accelerated in response to this incident.

Services Australia will continue to engage in substantial cyber security work, in line with the Government's risk-based approach, noting its systems include critical digital services and high-value data holdings. Government has invested around $160 million for cyber security uplift for Services Australia. This is in addition to work underway to update Services Australia's ICT architecture, which will further modernise and protect key ICT systems.

Action taken

- 2023 onward – ASD commenced publishing public advice on AI, and has regularly updated the Information Security Manual to include AI controls.
- May 2026 – Home Affairs issued guidance on cyber security readiness in the frontier AI era.
- June 2026 – ASD's ACSC and Five Eyes released a joint statement warning that frontier AI models would fundamentally transform offensive and defensive cyber capabilities and recommended actions for organisations to address this risk.
- September 2026 – Home Affairs issued PSPF direction to require entities to undertake legacy technology stocktake, risk management plan, and incorporate new risk reduction measures for Systems of Government Significance (in response to AI's impacts on the threat environment).

Recommendation

18. Commonwealth agencies engage with ASD to identify and assess opportunities to responsibly use AI-enabled cyber defence tools and techniques to enhance cyber security monitoring, penetration testing, threat detection and incident response capabilities.

References

[i] Prime Minister of Australia, 'Press conference – New York', 24 September 2026.
[ii] OpenAI, 'How we will do better for Australia', *OpenAI blog*, 28 September 2026.
[iii] OpenAI, 'How we will do better for Australia', *OpenAI blog*, 28 September 2026.
[iv] OpenAI, 'How we will do better for Australia', *OpenAI blog*, 28 September 2026.
[v] METR and Redwood Research, *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI/Hugging Face hacking incident*, METR and Redwood Research, 26 August 2026; United Nations Independent Scientific Panel on AI, *AI agents, misalignment and the risk of losing human control: Evidence from the OpenAI-Hugging Face Incident*, United Nations, September 2026 [advance unedited version].
[vi] Services Australia, *Report a cyber security system risk*, Services Australia website, 5 June 2026.
[vii] Department of Home Affairs, *Australian Government Protective Security Policy Framework: Release 2026*, Australian Government, 1 July 2026.
[viii] Reuters, OpenAI alerts more than 100 groups about rogue AI agent activity, *Reuters*, 2 October 2026; Cable et al,. *Early rogue AI agent activity and attempts to hack found on urlquery.net* , Transluce, 23 September 2026.
[ix] OpenAI, 'How we will do better for Australia', *OpenAI blog*, 28 September 2026.
[x] Department of the Prime Minister and Cabinet, *Rapid review into Australian government arrangements for an AI-driven cyber incident*, PM&C website, 24 September 2026.
[xi] Australian Signals Directorate, *Five Eyes cyber security agencies statement* [joint statement], Five Eyes, 22 June 2026.
[xii] United Nations Independent Scientific Panel on AI, *AI agents, misalignment and the risk of losing human control: Evidence from the OpenAI-Hugging Face Incident*, United Nations, September 2026 [advance unedited version].
[xiii] United National Independent International Scientific Panel on AI, *Thematic Brief on AI Agents, Misalignment and the Risk of Losing Human Control*, United Nations website, n.d.
[xiv] United Nations Independent Scientific Panel on AI, *AI agents, misalignment and the risk of losing human control: Evidence from the OpenAI-Hugging Face Incident*, United Nations, September 2026 [advance unedited version].
[xv] Wilson, C, 'Australia needs an AI 'early warning system', top cybersecurity chief warns', *ABC News*, 15 September 2026.
[xvi] Joint Select Committee on Artificial Intelligence, *Tuesday 6 October 2026*, 35.
[xvii] OpenAI, 'How we will do better for Australia', *OpenAI blog*, 28 September 2026.
[xviii] Australian Signals Directorate, *Five Eyes cyber security agencies statement* [joint statement], Five Eyes, 22 June 2026.
[xix] Department of Home Affairs, *Cyber coordinator*, Home Affairs website, n.d.
[xx] Australian Signals Directorate, *Five Eyes cyber security agencies statement* [joint statement], Five Eyes, 22 June 2026.
[xxi] Australian Signals Directorate Australian Cyber Security Centre, *The Commonwealth Cyber Security Posture in 2025*, Australian Signals Directorate, 12 February 2026.
[xxii] Australian Signals Directorate Australian Cyber Security Centre, *The Commonwealth Cyber Security Posture in 2024*, Australian Signals Directorate, 5 December 2024.
[xxiii] Department of Home Affairs, *2023-2030 Australian Cyber Security Strategy*, Department of Home Affairs, 2023.
[xxiv] Department of Home Affairs, *2023-2030 Australian Cyber Security Strategy Horizon 2 Action Plan*, Department of Home Affairs, 2026.
[xxv] Australian Cyber Security Centre, *Opportunities for AI in cyber defence*, Australian Signals Directorate, 8 September 2026.

Appendix A – Terms of reference

Purpose

These terms of reference are to guide a Department of the Prime Minister and Cabinet-led rapid-review of an artificial intelligence (AI) related cyber-incident affecting Australian Government systems. The objective of the review is to determine whether existing legislative, governance, and information-sharing arrangements are fit for purpose to prepare for, and respond to, a cyber incident involving AI. The review will also inform how to build and maintain resilient systems in the AI era.

Context

This Review follows reporting from OpenAI to the Australian Government of a non-public OpenAI model undertaking mis-aligned activity during an internet research task, resulting in a cyber incident. The review is occurring in the context of, and to inform, Australian Government preparedness for AI-driven cyber threats, the development of Australia's AI Standards, Australia's engagement on a global approach to common standards for safety incident reporting, and review of other Australian Government legislative, regulatory and crisis management arrangements.

The Department of the Prime Minister and Cabinet will undertake the review of Australian Government preparedness and response to the incident, in collaboration with the National Cyber Security Coordinator, ASD, AISI and Services Australia. This will inform recommendations on:

- Reporting requirements relating to AI-driven cyber-incidents, AI-identified cyber vulnerabilities and cyber-related AI safety incidents, including reporting obligations, thresholds, pathways, and systems.
- Commonwealth governance and information-sharing arrangements for managing incidents involving AI, including roles and responsibilities and escalation pathways.
- Engagement and information-sharing obligations of AI firms, including notification requirements and cooperation arrangements during incidents.
- The adequacy of existing system and legislative, regulatory and enforcement frameworks in deterring AI-driven cyber incidents, including whether current offences, liabilities, penalties and enforcement mechanisms are sufficient and effective.
- Mechanisms to strengthen networks and systems of government departments and agencies against AI vulnerabilities.

Appendix B – Review methodology

The Review was drafted based on desktop research, interviews and input from Review agencies and other relevant Commonwealth departments.

During the Review, interviews were conducted with:

- Attorney-General's Department
- Australian Cyber Security Centre, Australian Signals Directorate
- Australian Federal Police
- Commonwealth Scientific and Industrial Research Organisation
- Department of Industry, Science and Resources, including the AI Safety Institute
- Department of Foreign Affairs and Trade
- The National Cyber Security Coordinator and the National Office of Cyber Security, Department of Home Affairs
- Office of AI, Department of the Prime Minister and Cabinet
- Services Australia

Formal requests for information and/or written input into the Review was sought from:

- Attorney-General's Department
- Australian Signals Directorate, including the Australian Cyber Security Centre
- Australian Federal Police
- Department of Industry, Science and Resources, including the AI Safety Institute
- Department of Finance
- Department of Foreign Affairs and Trade
- Department of Home Affairs, including the National Office of Cyber Security
- Department of the Prime Minister and Cabinet, including the Office of AI
- Office of National Intelligence
- Services Australia
- The Treasury

Additional specific documentation was also sought from:

- Department of Agriculture, Fisheries and Forestry
- Department of Climate Change, Energy, the Environment and Water
- Department of Defence
- Department of Education
- Department of Employment and Workplace Relations
- Department of Health, Disability and Ageing
- Department of Infrastructure, Transport, Regional Development, Communications, Sport and the Arts
- Department of Social Services
- Department of Veterans' Affairs

Appendix C – AGCMF crisis responsibilities by hazard

The Australian Government Crisis Management Framework (AGCMF) outlines designated Lead Ministers, Lead Coordinating Senior Officials and Australian Government Coordinating Agencies for identified hazards.

The following summarises those arrangements. While this Review is primarily concerned with a cyber incident, crisis arrangements for all-hazards are outlined, noting the cross-cutting potential risks of harmful AI activity across multiple domains.

Cyber Incidents	Domestic biosecurity crises	Domestic energy supply crisis	Domestic natural hazard disasters	Maritime terrorist incidents[13]
Lead Minister: Minister responsible for Cyber Security **AUSGOV Coordinating Agency:** Home Affairs **Lead Coordinating Senior Official:** National Cyber Security Coordinator or other relevant Deputy Secretary, Home Affairs	**Lead Minister:** Minister responsible for Agriculture **AUSGOV Coordinating Agency:** Department of Agriculture, Fisheries and Forestry (DAFF) **Lead Coordinating Senior Official:** Deputy Secretary responsible for Biosecurity, DAFF	**Lead Minister:** Minister responsible for Energy **AUSGOV Coordinating Agency:** Department of Climate Change, Energy, the Environment and Water (DCCEEW) **Lead Coordinating Senior Official:** Deputy Secretary responsible for Energy, DCCEEW	**Lead Minister:** Minister responsible for Emergency Management **AUSGOV Coordinating Agency:** National Emergency Management Agency (NEMA) **Lead Coordinating Senior Official:** Deputy Coordinator General, Emergency Management and Response, NEMA	**Lead Minister:** Minister responsible for Home Affairs **AUSGOV Coordinating Agency:** Australian Border Force (Maritime Border Command) **Lead Coordinating Senior Official:** Commander Maritime Border Command

[13] Within the Australian Security Forces Authority Area.

Domestic terrorist incidents	Incidents involving an offshore petroleum facility[14]	Transport incidents within Australia[15]	Domestic public health crises	International crises
Lead Minister: Minister responsible for Home Affairs **AUSGOV Coordinating Agency:** Home Affairs **Lead Coordinating Senior Official:** Commonwealth Counter-Terrorism Coordinator	**Lead Minister:** Minister responsible for Resources **AUSGOV Coordinating Agency:** DISR **Lead Coordinating Senior Official:** Deputy Secretary responsible for Resources, DISR	**Lead Minister:** Minister responsible for Transport **AUSGOV Coordinating Agency:** Department of Infrastructure, Transport, Regional Development, Communications, Sport and the Arts (DITRDCSA) **Lead Coordinating Senior Official:** Deputy Secretary responsible for Transport, DITRDCSA	**Lead Minister:** Minister responsible for Health **AUSGOV Coordinating Agency:** Department of Health, Disability and Ageing **Lead Coordinating Senior Official:** Chief Medical Officer, Department of Health, Disability and Ageing	**Lead Minister:** Minister responsible for Foreign Affairs **AUSGOV Coordinating Agency:** DFAT **Lead Coordinating Senior Official:** Deputy Secretary responsible for Consular and Crisis Management, DFAT

Domestic security-related incidents (excl. terrorist incidents)	Radiological / nuclear incidents	Space weather events	Novel or ambiguous hazard
Lead Minister: Minister responsible for Home Affairs **AUSGOV Coordinating Agency:** Home Affairs **Lead Coordinating Senior Official:** Head of National Security, Home Affairs	**Lead Minister:** Minister responsible for Health and Aged Care **AUSGOV Coordinating Agency:** Australian Radiation Protection and Nuclear Safety Agency (ARPANSA) **Lead Coordinating Senior Official:** Chief Executive Officer, ARPANSA	**Lead Minister:** Minister responsible for Emergency Management **AUSGOV Coordinating Agency:** National Emergency Management Agency (NEMA) **Lead Coordinating Senior Official:** Deputy Coordinator General, Emergency Management and Response, NEMA	**Lead Minister:** Minister responsible for Emergency Management **AUSGOV Coordinating Agency:** NEMA **Lead Coordinating Senior Official:** Deputy Coordinator General, Emergency Management and Response, NEMA

[14] In Commonwealth waters.

[15] This includes the Australian Search and Rescue Region and the Australian Exclusive Economic Zone.

Appendix D – Glossary

Term	Definition
ACSC	Australian Cyber Security Centre
AFP	Australian Federal Police
AGCMF	Australian Government Crisis Management Framework
AGD	Attorney-General's Department
AI	Artificial Intelligence
AISI	Australia's AI Safety Institute
APEC	Asia-Pacific Economic Cooperation
ARPANSA	Australian Radiation Protection and Nuclear Safety Agency
ASD	Australian Signals Directorate
ASEAN	Association of Southeast Asian Nations
DFAT	Department of Foreign Affairs and Trade
DISR	Department of Industry, Science and Resources
DITRDCSA	Department of Infrastructure, Transport, Regional Development, Communications, Sport and the Arts
Enabling agency	As defined in the Australian Government Crisis Management Framework, an enabling agency is an Australian Government agency that administers relevant programs, provides specialist technical, scientific, intelligence or information capabilities or conducts any other enabling activities to support consequence management activities. A key example in relation to AI is the AI Safety Institute.
NEMA	National Emergency Management Agency
NOCS	National Office of Cyber Security
PM&C	Department of the Prime Minister and Cabinet
PSPF	Protective Security Policy Framework
UN	United Nations