



Submission to the Office of AI's Consultation on Australian Standards for AI

9 October 2026

Setting the Standard: An Australian Model for AI Incident Monitoring & Reporting

Introduction and Executive Summary

The Australian Standards for Artificial Intelligence (AI) are being developed at a critical time. AI capabilities are advancing rapidly and are creating enormous opportunities, while also presenting escalating risks. The Australian Standards must enable Australia to seize the benefits of advanced AI, by effectively managing its risks.

[Global Shield Australia](https://globalshieldpolicy.org) is an independent, non-profit policy advocacy organisation working to reduce Australia's vulnerability to catastrophic risk. Our work covers all hazards, including a particular focus on AI. This submission focuses on the Consultation Paper's questions regarding the security, safety, and transparency of AI training, as well as the scope of the Australian Standards.

Global Shield Australia's work has had a particular focus on AI incident monitoring and reporting (AIIMR). **Appendix A** of this submission sets out a detailed proposal covering the key elements that an Australian AIIMR regime should contain. This is informed by our extensive research and monitoring of existing AIIMR systems overseas, incident reporting obligations in other sectors, and engagement with experts.

This submission also sets out four other ways in which the Australian Standards could reduce the risk from frontier AI. First, by obliging AI companies to effectively manage risk. Second, through transparency requirements in relation to training data and processes. Third, by ensuring Australia's whistleblower protections properly apply to frontier AI company staff. And fourth, by setting clear requirements in relation to independent evaluation and verification of safety and security practices.

This submission also emphasises that, to be effective, the Australian Standards for AI cannot be restricted only to AI data centres and AI training. The Standards should apply throughout the AI lifecycle, including to inference, and to all models served to Australian users regardless of where they are trained.

The Australian Standards for AI must ensure that Australia has the governance frameworks necessary to successfully navigate our AI future. They should enable Australia to manage AI risk, secure its benefits, and offer a model to the region and the world, setting the standard for others to follow.

Summary of recommendations

Recommendation 1: The Australian Standards for AI should apply not only as a condition of training frontier AI models, but also to the evaluation, internal use, and deployment of frontier AI models in Australia and when models are served to Australians from overseas.

Recommendation 2: The Australian Standards for AI should establish, or be accompanied by, a mandatory AI incident monitoring and reporting regime, with the elements set out in **Appendix A**, including: a duty to monitor; reportable incidents defined to include near misses and control failures; tiered reporting deadlines; and an agency empowered to act upon reports.

Recommendation 3: Drawing on best practices overseas, the Australian Standards for AI should: require AI companies and AI data centres to adopt effective risk management frameworks; require transparency regarding training data and processes; ensure whistleblower protections apply to, and are appropriate for, frontier AI company staff; and build on industry moves towards independent verification by requiring such verification in Australia and investing in research and development on verification technologies.

Recommendation 4: The legislation containing the Australian Standards for AI should establish the regulatory framework and set core obligations, while specific details such as thresholds for coverage should be set or able to be updated through delegated instruments. The Standards should also be subject to regular review.

We thank the Office of AI for its important work and consideration of our submission. We would welcome the opportunity to discuss our recommendations further.

Contact: australia@globalshieldpolicy.org

1. The Australian Standards for AI must apply across the entire AI lifecycle

Getting the scope of the Australian Standards for AI correct will be key to their effectiveness. The Consultation Paper refers to the Standards applying to frontier AI training in Australia. However, the Paper does not specify whether they will also apply to internal deployment and testing of frontier AI models, or the serving of such models to Australian users or from Australia to other jurisdictions (i.e. inference).¹

While frontier AI model training has its own set of specific risks that need to be managed, many of the most serious concerns arising from frontier AI also apply once a model is trained and in deployment. These include misuse of models by malicious actors and dangerous interactions of AI systems with other software and information technologies (including other AI systems).² Indeed, the recent incident involving Services Australia would not have been covered by a regime that only focused on ‘training’, if that term was too narrowly defined.³

Other jurisdictions have not limited the scope of their AI regulations to just ‘training’. For example, legislation in California and New York does not apply to just training of frontier AI models.⁴ The European Union’s (EU) AI Act covers development and deployment.⁵ South Korea and Vietnam also apply their AI Acts beyond just AI training.⁶

Similarly, the Australian Standards need to clearly capture all phases of the AI model lifecycle, including after a model is deployed to the market.

Recommendation 1: The Australian Standards for AI should apply not only as a condition of training frontier AI models, but also to the evaluation, internal use, and deployment of frontier AI models in Australia and when models are served to Australians from overseas.

¹ See Department of the Prime Minister and Cabinet, [Getting it right: Building AI infrastructure that works for Australia](#) (17 Sep 2026) (Consultation Paper) p. 26: “AI training: AI training builds an AI model using large volumes of data and significant computing power. It includes pre-training, which establishes the model’s core capabilities, and post-training, which refines how it performs. In simple terms, training is how AI is built.”

² See, e.g., L. Hammond et al., [Multi-Agent Risks from Advanced AI](#) (Cooperative AI Foundation, Feb 2025); METR, [Brief independent investigation of agents’ behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident](#) (26 Aug 2026); Gradient Institute, [When single AI agents become a swarm](#) (28 Sep 2026).

³ OpenAI referred to misaligned activity “during training” but also that the Services Australia incident occurred during an “internal evaluation”, see ABC News, [OpenAI hacked Medicare portal. Prime Minister Anthony Albanese says](#) (24 Sep 2026).

⁴ *Transparency in Frontier Artificial Intelligence Act* (California), §22757.11(c), (e)(1) and (h); *Responsible AI Safety and Education Act* (New York), §1425.

⁵ AI Act (EU), art 2(1)(a)-(c).

⁶ AI Basic Act (Korea), arts 2(4), 32, and 34; *Law on Artificial Intelligence* (Vietnam), art 12(1).

2. An Australian model for mandatory AI incident monitoring and reporting

The Consultation Paper rightly recognises the value of requiring reporting of “*AI incidents to relevant Australian authorities*”.⁷ A properly designed mandatory AI incident monitoring and reporting (AIIMR) regime will provide the government critical intelligence on AI risk, enable appropriate response to harms and threats, and empower Australians harmed by AI.

There are several models overseas and in other sectors that can be drawn from when developing Australia’s AIIMR regime. However, the Australian Standards for AI must not just duplicate systems established elsewhere. Elements of AIIMR systems overseas have been shown to be insufficient in light of recent AI incidents and would not meet the expectations of Australians.⁸

Global Shield Australia has prepared a detailed proposal setting out the key elements needed for an effective and world leading Australian AIIMR regime (see **Appendix A**). This work is informed by practice overseas, lessons learned since those systems were implemented, existing non-AI incident reporting schemes within Australia, and expert analysis and commentary. Instituting a properly designed AIIMR regime in Australia can both help protect Australians and establish a precedent for our region and more broadly.

The key elements of the proposed approach are set out in **Table 1**.

Recommendation 2: The Australian Standards for AI should establish or be accompanied by a mandatory AI incident monitoring and reporting regime, with the elements set out in **Appendix A**, including: a duty to monitor; reportable incidents defined to include near misses and control failures; tiered reporting deadlines; and an agency empowered to act upon reports.

⁷ Consultation Paper, p. 21.

⁸ See, e.g., Z. Kara, [Improving Frontier AI Incident Reporting Regimes](#) (GovAI, Policy Brief, September 2026).

Table 1. Seven key elements of an effective AIIMR regime
1. The AIIMR must have a clear purpose to inform its design and implementation

The aims of Australia's AIIMR regime should be to:

- (a) Activate incident and emergency response mechanisms to minimise harms and contain incidents;
- (b) Identify factors that contribute to incidents;
- (c) Ensure appropriate corrective and remediation action is undertaken; and
- (d) Enable learning and improved risk controls.

2. Monitoring and reporting must be mandatory for covered entities, with avenues for voluntary reporting by those impacted by an AI incident

Covered entities must be required to have effective monitoring systems in place to ensure they are proactively identifying incidents. The length of time taken to identify and report AI incidents occurring in early 2026 demonstrates the need for AI companies to have more robust incident monitoring systems in place. For example, following the reporting of the Hugging Face incident in July 2026, several AI companies commenced internal reviews and identified a range of other incidents, stretching back potentially as far as November 2025.⁹

Australia's AIIMR regime must require comprehensive and effective monitoring to avoid this happening again. The regime should specify what records need to be preserved to enable proper analysis of incidents when they happen,¹⁰ and ensure that actions can be attributed to relevant AI agents.¹¹

Voluntary reporting pathways should also be available for users of AI systems, those impacted by incidents, and whistleblowers.¹²

3. Monitoring and reporting obligations must apply to all entities best placed to detect incidents

The AIIMR obligations should apply to at least three categories of entities: (1) companies training and serving frontier AI models (identified by compute used, assessed capabilities,

⁹ See, e.g., J. Taylor, [OpenAI has 'work to do to rebuild trust' in Australia, executive tells AI inquiry](#) (the Guardian, 6 Oct 2026); Anthropic, [Investigating three real-world incidents in our cybersecurity evaluations](#) (30 Jul 2026); OpenAI, [OpenAI and Hugging Face partner to address security incident during model evaluation](#) (21 Jul 2026); Centre for Long-Term Resilience, [AI loss of control incidents are worsening](#) (29 Aug 2026).

¹⁰ See Z. Kara, [Improving Frontier AI Incident Reporting Regimes](#) (GovAI, Policy Brief, September 2026), p. 13 for list of the evidence that AI companies should be required to preserve in relation to AI incidents.

¹¹ Ibid 10-12.

¹² See, e.g., *Transparency in Frontier Artificial Intelligence Act* (California), §22757.13(a) and (d); *Transport Safety Investigation (Voluntary and Confidential Reporting Scheme) Regulation 2012* (Cth), reg 7; *Labor Code* (California), §1107.1.

and risk), (2) data centre operators hosting the training or inference of such models, and (3) third parties hosting or testing such models.¹³

Each of these entities will see different aspects or types of potential incidents. But each can contribute important evidence that others may lack to create a fuller picture of the threat landscape. Multiple reporting sources can also help to keep all covered entities accountable regarding their compliance with their AIIMR obligations.

4. Reportable incidents need to include serious incidents and harms (defined to ensure all material incidents are captured) as well as near misses

While not all incidents and harms need to be reported, it is critical that the threshold for reporting is not set unrealistically high. Some overseas regimes adopt too high a threshold for what is reportable (such as requiring 50 deaths, one billion dollars of damage, or a small number of specific model failures where these increase the risk of such harm or result in death or injury). The aim for the AIIMR regime should be to avoid the most significant harms from occurring by requiring reporting of lesser incidents and near misses.

As such, an Australian AIIMR should be calibrated against similar regimes in other industries, such as aviation, where a single death or serious harm is required to be reported.¹⁴ Incidents where harm is narrowly avoided should also be captured, as should AI-specific novel occurrences demonstrating new or increased risks.

Reportable incidents under Australia's AIIMR should be defined with sufficient specificity, supported by clear guidance, across three categories:

- (a) **Harms:** where a frontier AI model, as a result of misuse, malfunction or autonomous action, causes, materially contributes to, or is reasonably likely to have caused or contributed to:
 - (i) One or more deaths or serious injuries;
 - (ii) a serious criminal offence (including in relation to chemical, biological, radiological, nuclear, or offensive cyber capabilities);
 - (iii) a serious cyber security incident,¹⁵ particularly involving government systems;
 - (iv) disruption to critical infrastructure;
 - (v) serious damage to property or economic interests
 - (vi) serious harm to the environment; or
 - (vii) prejudice to Australia's national security.
- (b) **Model, oversight, and security failures (regardless of whether harm results):**
 - (i) loss of control of a frontier AI model;

¹³ Noting that some data centre operators may have existing cyber security incident reporting obligations under the *Security of Critical Infrastructure Act (SOCI Act)*.

¹⁴ See, e.g., *Transport Safety Investigation Regulations 2021* (Cth), regs 6, 6A, and 11A-11C; *Work Health and Safety Act 2011* (Cth) ss 35–36; *Therapeutic Goods (Medical Devices) Regulations 2002* (Cth) reg 5.7.

¹⁵ Defined, e.g., as in SOCI Act, s 12M.

- (ii) a model subverts or evades, or credibly attempts to subvert or evade, monitoring or controls, or acts materially outside of or contrary to its instructions;
 - (iii) unauthorised access to or exfiltration of a model's weights, or credible attempts at either (including by the model itself);
 - (iv) insider threats such as employee espionage or misuse of or tampering with models and internal data; and
 - (v) other serious cyber security incidents targeting a model or its host.
- (c) **Near misses:** where any of the above incidents are avoided by an unplanned intervention or chance. This is similar to aviation and workplace safety reporting regimes where events that could have escalated to actual harm are reportable.¹⁶

Further guidance on thresholds and definitions will be needed for many of the categories above to ensure clarity for industry. Some exceptions for testing and evaluations may also be necessary. However, these should be narrowly drafted such that they only cover behaviour that was explicitly and intentionally elicited from a model, and where that behaviour remained within the control and containment of the evaluator.

5. Reporting deadlines should be tiered by severity, with initial reports followed by comprehensive and then final reports for serious incidents

Timelines for reporting should commence when the entity is aware of the facts supporting a reasonable belief that a reportable incident has occurred. Deadlines for initial reports should be tiered by severity:

- (a) 24 hours where there is an imminent risk of, or actual, death or serious injury;
- (b) 48 hours for a widespread or systemic risk; and
- (c) otherwise 72 hours.¹⁷

Initial reports should contain what is minimally known at the time of the report. They should be followed by a comprehensive report containing information on the causes of the incident, its impact, and responses. A final report, following a full investigation where necessary, should set out the lessons learned and measures implemented to prevent future incidents.

6. Incident reports must go to an agency empowered to assess, share and act on them

Reports should go to an agency that has the technical capability to assess and act upon the full range of potential AI incidents.¹⁸ If given statutory backing and the necessary resources, Australia's AI Safety Institute (**AISI**) would likely be the best placed agency for this role.

¹⁶ *Transport Safety Investigation Regulations 2021* (Cth), reg 6A; *Work Health and Safety Act 2011* (Cth) s 37.

¹⁷ See, e.g., *SOCI Act ss 30BC-30BD*; *Responsible AI Safety and Education Act* (New York), §1422(3) (New York); *Transparency in Frontier Artificial Intelligence Act* (California), §22757.13(c); *AI Act* (EU), art 73(2)-(5).

¹⁸ In the EU reports go to market surveillance authorities (*AI Act* (EU), arts 55(1)(c) and 73(1)); in California to a section of the Office of Emergency Services (*Transparency in Frontier Artificial Intelligence Act* (California), §22757.13); and in New York to a section of the Department of Financial Services (*Responsible AI Safety and Education Act* (New York), §1420(16)).

The Australian Signals Directorate (**ASD**) is also an alternative option. However, its role within the intelligence community may limit its ability to engage publicly and share information. Its particular expertise in cyber may also be less relevant to other forms of AI incidents (such as those involving biological threats).

The AIIMR agency should be required to pass incident reports to relevant regulators for response. For example, ASD for cyber incidents, the Australian Centre for Disease Control for human health and biological incidents, and the Cyber and Infrastructure Security Centre for critical infrastructure issues. This should also allow for deemed compliance with equivalent reporting obligations under other regulatory frameworks, as long as other agencies also refer relevant incident reports to the AIIMR agency for compilation.

The agency should also have powers to investigate reports, or order independent investigation. It should have the power to compel further information from reporting entities, obtain access to the relevant model, and issue warnings or advisory notices to the industry or public. It should also issue regular reporting on the overall AI risk environment and how safety and security practices can be improved. The agency should also have powers to require entities to prepare corrective action plans following incidents, and to direct that an AI model or system be restricted, modified, or withdrawn from use in Australia.¹⁹

7. The AIIMR regime must be adaptable to keep pace with technological developments; it should also be interoperable with overseas regimes without being limited by them

To ensure the AIIMR regime can keep pace with the technology, the relevant legislation should put in place the framework and core definitions to allow the regime to be established. However, key thresholds and definitions (such as for the meaning of ‘frontier AI’) should be able to be updated through delegated instruments. This will allow these to be reviewed and updated regularly as the technology and threat environment changes.

Work with international partners on incident reporting and response will also be key. The government should establish incident and information sharing channels with trusted partners, and also support our near neighbours with incident responses.

To the extent possible, Australia’s requirements for format or other procedural aspects of reports should be aligned and interoperable with regimes overseas. This can be achieved by using a baseline such as the OECD’s Common Reporting Framework and by considering commonalities across the EU, US, and emerging regimes in Vietnam and South Korea.²⁰ However, interoperability does not mean limiting what incidents need to be reported, the contents of such reports, nor the powers of agencies to respond.

¹⁹ See, e.g., *AI Act* (EU), arts 91-93; *Transparency in Frontier Artificial Intelligence Act* (California), §22757.13. In Australia, see, e.g., *Australian Consumer Law* (Cth), ss 122 (compulsory recall), 129 (safety warning notices); *Therapeutic Goods Act 1989* (Cth) ss Part 3-2, Div 2A, Part 3-2A, Div 8, and Part 4-9.

²⁰ OECD, [Towards a Common Reporting Framework for AI Incidents](#) (Feb 2025).

Box 1. How an AIIMR would have applied to the Services Australia OpenAI incident

On 18 June 2026, an unreleased OpenAI agent undergoing testing circumvented blocks on Services Australia's Medicare Statistics Reporting Service website and accessed non-public information. OpenAI discovered the incident in August and reported it to the Australian Government on 10 September via a public Services Australia email address. ASD was informed of the incident on 15 September, and the broader public on 24 September.²¹

If Australia had an AIIMR of the kind described in this submission:

- (a) **OpenAI would likely have identified the incident sooner:** OpenAI would have been required to effectively log and monitor the actions of its agents (including during testing), making it more likely that the incident would have been identified in June, not months later in August.
- (b) **OpenAI would have been legally obliged to report the incident:** the incident would likely have been reportable on two bases. First, as the agent acted beyond its tasking and breached OpenAI's guardrails. Second, as it appears to have gained unauthorised access to an Australian government system.
- (c) **The government would have been notified within 72 hours:** OpenAI would have been required to provide an initial report of the incident within 72 hours. This would have gone through a formal reporting channel to the AISI, which could then refer the matter to ASD and Services Australia.
- (d) **The government would have had powers to investigate and issue warnings to others:** the AISI could have requested preserved logs relating to the incident, ordered a more detailed investigation, and notified other agencies (or had ASD notified other agencies) of the need to address similar issues in their publicly facing services.
- (e) **OpenAI would have been obliged to provide a full explanation for what went wrong and what corrective action had been taken:** comprehensive and final reports would have followed OpenAI's initial report. These would set out the company's analysis of the causes of the incident, how the company responded, and what actions were taken to prevent recurrence of the same issue in the future.
- (f) **There would be broader learning:** de-identified lessons from the incident could be published to improve the safety and security of the broader cyber and AI ecosystem.

This would have improved the response in this specific case. It would also be all the more important where the harm is far more severe, such as if an AI system is being used to plan a terrorist attack or for biological weapons development. In such cases, it is crucial that the government is promptly informed, has the powers to investigate and respond, and then has mechanisms in place to ensure lessons learned are quickly promulgated.

²¹ ABC News, [OpenAI hacked Medicare portal. Prime Minister Anthony Albanese says](#) (24 Sep 2026); OpenAI, [How We Will Do Better for Australia](#) (28 Sep 2026).

3. Additional safety and security measures for frontier AI companies and data centres

In addition to AIIMR, the Australian Standards for AI should also incorporate measures adopted in other AI regulatory regimes in place overseas. Key aspects of those regimes that Australia should adopt are:

1. **Risk management obligations.** This includes obligations to publish safety and security frameworks detailing how AI companies assess risk, apply mitigations, and respond to safety and security incidents.

Models for these obligations exist in regimes in the EU, California, New York, Illinois, South Korea and Vietnam.²² Australia's *Security of Critical Infrastructure Act (SOCI Act)* Critical Infrastructure Risk Management Program (**CIRMP**) requirement is a further precedent for this kind of obligation.²³ For example, frontier AI data centres should be addressing at least the risks specified in **Table 2**.

Table 2. Examples of frontier AI data centres risks	
Risk	Example
Exfiltration of frontier model weights	RAND has identified 38 attack vectors for stealing high-value model weights, from insiders to supply-chain compromise. ²⁴
Misuse of Australian-hosted compute by malicious actors	Anthropic has reported the use of Claude Code by state-sponsored groups for espionage. ²⁵
Loss of control during training, evaluation or deployment	AI agents have breached sandboxes during training and evaluation. ²⁶ Increasingly capable models raise concerns that autonomous AI systems could operate without human supervision. ²⁷
Synchronised training loads destabilising the grid	Synchronised AI training loads can create power swings within short periods of time. If these resonate with grid frequencies, they can potentially damage equipment or cause blackouts. ²⁸ Synchronised disconnection in response to grid faults creates similar risks. ²⁹

²² AI Act (EU), art 55; *Transparency in Frontier Artificial Intelligence Act* (California), §22757.12; *Responsible AI Safety and Education Act* (New York), §1421; *AI Basic Act* (Korea), art 32; *Law on Artificial Intelligence* (Vietnam), arts 13, 14.

²³ SOCI Act Part 2A.

²⁴ RAND, [Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models](#) (21 Nov 2024).

²⁵ Anthropic, [Disrupting the first reported AI-orchestrated cyber espionage campaign](#) (13 Nov 2025).

²⁶ METR, [Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI/Hugging Face hacking incident](#) (26 Aug 2026).

²⁷ D. Ball, [On the Loose](#) (Hyperdimensional, 1 Sep 2026).

²⁸ E. Choukse et al., [Power Stabilization for AI Training Datacenters](#) (21 Aug 2025); SemiAnalysis, [AI Training Load Fluctuations at Gigawatt-scale – Risk of Power Grid Blackout?](#) (26 Jun 2025).

²⁹ North American Electric Reliability Corporation, [Incident Review: Considering Simultaneous Voltage-Sensitive Load Reductions](#) (8 Jan 2025).

2. **Transparency.** In addition to the transparency mandated under an AIIMR regime, the Australian Standards for AI should also apply transparency obligations to AI companies in relation to their training data and methods. These should build on existing models, such as those in California and the EU, particularly to ensure appropriate records are maintained and are accessible by a regulator or investigator when needed. To that end, Global Shield Australia supports the submission of the Human Technology Institute to this consultation.³⁰
3. **Whistleblower protections.** Frontier AI company staff are best placed to know and assess when safety and security practices are insufficient or when a model is not performing appropriately. Australia has existing whistleblower protections regarding misconduct and breaches of the law.³¹ However, it is unclear whether these would be sufficient to cover the kind of harms and incidents an AI company staff member may need to report (i.e. where there is a danger to the public but no law has been broken).

As part of the development of the Australian Standards for AI, the government should review existing whistleblower protections, particularly against practice in California and elsewhere,³² to identify where additional protection is needed to support disclosure of specific and substantial AI-related dangers to the public.

4. **Independent verification and verification technologies.** Major AI companies are moving to embed independent evaluators within their processes, providing access to pre-release models, evaluations, and the right to publish their analysis.³³

Australia should leverage this work by requiring AI companies to host and fund independent verification organisations (IVOs) in Australia to assess the capabilities and risks associated with models, and certify safety and security practices or guardrails.³⁴ This would provide additional assurance as to the quality of the safety and security practices of AI companies; it would also help develop a local AI sector and expertise within Australia. This would also require strong protections against conflicts of interest and measures to ensure the independence of the IVOs.

For data centres, Australia could also mandate the use of hardware verification mechanisms, or investment in this technology, to provide assurance regarding how compute assets are being used.³⁵ Development of this technology would also contribute to international efforts around AI security, by enabling verification of compliance with future treaty commitments.

³⁰ Human Technology Institute, *Australian Standards for AI and Training Data Transparency* (9 Oct 2026).

³¹ *Corporations Act 2001* (Cth) Part 9.4AAA; *Public Interest Disclosure Act 2013* (Cth).

³² *Labor Code* (California), §1107.1; *AI Act* (EU), art 87 (EU); and *Artificial Intelligence Safety Measures Act* (Illinois), s 20.

³³ D. Amodei, [We Must Pace the Frontier](#) (Sep 2026); OpenAI, [Priorities and Principles for Effective Third Party Assessments](#) (22 Sep 2026).

³⁴ See, generally, Fathom, [Independent Verification Organizations](#).

³⁵ See, e.g., Z. Richardson, [Middle Powers Should Invest in AI Verification](#) (AI Policy Bulletin, 24 Sep 2026).

Recommendation 3: Drawing on best practices overseas, the Australian Standards for AI should: require AI companies and AI data centres to adopt effective risk management frameworks; require transparency regarding training data and processes; ensure whistleblower protections apply to, and are appropriate for, frontier AI company staff; and build on industry moves towards independent verification by requiring such verification in Australia and investing in research and development on verification technologies.

4. Implementation

The Australian Standards for AI must be able to keep pace with rapid technological developments. The legislation instituting the Standards should therefore focus on their core elements, including duties and responsibilities, while leaving implementation details to delegated rules or Ministerial designation (for example, thresholds, contents of incident reports, or recognition of equivalent regimes). The Standards should also be subject to mandatory regular reviews. This approach has precedent in Australian regimes such as the SOCI Act and *Biosecurity Act*, and also overseas.³⁶

Recommendation 4: The legislation containing the Australian Standards for AI should establish the regulatory framework and set core obligations, while specific details such as thresholds for coverage should be set or able to be updated through delegated instruments. The Standards should also be subject to regular review.

Conclusion

The first nine months of 2026 have demonstrated the insufficiency of the current safety and security controls being applied by frontier AI companies. The Australian Standards for AI are Australia's opportunity to contribute to fixing this. By ensuring the Standards apply across the AI lifecycle, establishing a robust mandatory AIIMR regime, and drawing on and improving other elements of overseas practice, Australia will be able to successfully manage the risks of frontier AI while securing its benefits. If done right, we can also again set the standard for others to follow.

Global Shield Australia thanks the Office of AI for its consideration of this submission and would welcome the opportunity to brief officials further on our recommendations and work.

Contact: australia@globalshieldpolicy.org

³⁶ See, e.g., SOCI Act, Part 2A; *Biosecurity Act* 2015 (Cth) s 154; AI Act (EU), arts 51(3), 112.

Appendix A: An Australian AI Incident Monitoring and Reporting Regime

The Consultation Paper recognises that AI incident reporting is an important part of the regulatory toolkit that Australia needs to ensure the development of a safe and secure Australian AI sector. Getting the design of such a reporting regime right will be key, including ensuring all relevant incidents are captured and reported, that the government has appropriate powers to respond to and learn from incidents, and to build upon lessons learned from approaches taken overseas.

This Appendix sets out the key elements of an AI Incident Monitoring and Reporting (AIIMR) regime with recommendations for the approach Australia should take.

	Element & description	An Australian approach
1. Purpose		
1.1	Purpose of the regime <i>The AIIMR regime must have a clear purpose to ensure a coherent approach is taken to its various elements, such as what incidents are captured, how reports are dealt with, and what powers the receiving agency has to respond.</i>	An Australian AIIMR should have four purposes: (a) Activating incident and emergency response mechanisms to minimise harms and contain incidents; (b) Identifying factors that contribute to incidents; (c) Ensuring appropriate corrective and remediation action is undertaken; and (d) Enabling learning and improved risk controls.
	Precedents: Cyber Security Act 2024, s 3 (aims include to facilitate response to incidents, prevent and improve the detection of incidents, and facilitate sharing of information about incidents); Transport Safety Investigation Act 2003, s 12AA (functions include to identify factors that contribute to safety matters and publicly report on investigations); AI Act (EU), recital 155 (aims include to ensure AI providers can learn from experience to improve systems, designs, and development, and take timely response action).	
2. Scope – who and what is covered		
2.1	Covered AI models <i>Not all AI models pose the same level of risk. Limiting the AIIMR regime’s coverage will minimise regulatory burden and ensure the system is not overwhelmed.</i>	Frontier AI models with coverage determined by: (a) Compute used in their training (e.g. >10 ²⁶ FLOPs); (b) The breadth of their training data; (c) Their capabilities, generality, adaptability, and assessed level of risk; and/or (d) Ministerial designation. These thresholds should be set in regulations, not statute, to enable them to be regularly reviewed and updated.

	Element & description	An Australian approach
		The Consultation Paper contemplates applying incident reporting to frontier AI training. However, to be effective, this obligation should continue to apply to both internal use and external deployments. It must apply throughout the AI model's life cycle, especially when the product, service, or model is actively available in the Australian market.
	Precedents: <i>Transparency in Frontier Artificial Intelligence Act (California), §22757.11(f), (i), (j) (referring to compute, breadth of training data, generality of output, adaptability, and revenue); AI Act (EU), arts 3(63)-(65), 51 (referring to training data size, capabilities, and compute); AI Basic Act (Korea), s 32(1) (allowing compute thresholds to be set by Presidential Decree); Security of Critical Infrastructure Act, s 51 (enabling Ministers to declare assets to be critical infrastructure assets).</i>	
2.2	Entities required to monitor and report <i>Mandatory monitoring and reporting obligations should apply to those entities with access to and visibility of AI incidents and their impacts.</i>	Mandatory monitoring and reporting obligations should apply to: (a) Companies training or deploying frontier AI models in Australia or serving such models to Australian users; (b) Australian data centres hosting frontier AI training or inference, in regard to incidents they can observe (e.g. security incidents; unauthorised access/use); and (c) Third parties evaluating, training, or hosting frontier AI models. Pathways for voluntary reports should also be available from a broad range of stakeholders and users.
	Precedents: <i>Transparency in Frontier Artificial Intelligence Act (California), §22757.11(h), (j) (defining frontier developer); Responsible AI Safety and Education Act (New York), §1425 (specifying the Act applies to frontier models developed, deployed, or operating in New York); AI Act (EU), art 3(3), (4) (defining provider and deployer); Transport Safety Investigation (Voluntary and Confidential Reporting Scheme) Regulation 2012 (REPCON) (voluntary transport safety incident reporting regime); Security of Critical Infrastructure Act, Part 2B (cyber security incident notification requirements).</i> <i>See also C. Bullock et al., Legal Considerations for Defining "Frontier Model" (LawAI, Working Paper No. 2-2024, 2024); and M. Pistillo et al., The Role of Compute Thresholds for AI Governance, 1 George Washington Journal of Law and Technology 26 (2025).</i>	
2.3	Reportable incidents <i>The AIIMR regime should capture clearly defined harms where these have actually occurred or have been narrowly avoided. It should also apply to specific occurrences and events that indicate heightened risk.</i>	Reportable incidents should be clearly defined and objectively determinable events. A reportable incident should be defined as the occurrence of an event in which an AI model, agent, product, or service, whether through misuse or malfunction (including misaligned function), caused, materially contributed to, or is

	Element & description	An Australian approach
		reasonably likely to have caused or contributed to: (a) One or more deaths or serious injuries; (b) The commission of serious criminal offences, including development or deployment of chemical, biological, radiological, or nuclear weapons, or cyber attacks (whether by a human or AI); (c) A serious cyber security incident, particularly involving government systems; (d) Disruption to critical infrastructure; (e) Serious damage to property or economic interests (with defined thresholds); (f) Serious harm to the environment (with defined thresholds); or (g) Prejudice to Australia's national security (with clear guidance to industry). Reportable incidents should also include the following, whether or not harm results: (a) Loss of control of a frontier AI model; (b) A frontier AI model subverting, or credibly attempting to subvert, monitoring or controls, or acting materially outside of or contrary to its instructions (except when being evaluated to specifically test such behaviours, with stringent controls applied); (c) Unauthorised access to, or exfiltration of, frontier AI model weights, or a credible attempt to do so (whether by a human or AI); (d) Insider threats such as employee espionage or misuse of or tampering with models and internal data; and (e) Serious security and cyber incidents targeting a frontier AI model or host data centre. Reportable incidents should also include near misses, where any of the above harms or occurrences was avoided by an unplanned intervention or through chance.
		Precedents: <i>Transparency in Frontier Artificial Intelligence Act (California), §22757.11(c), (d) (definitions of</i>

Element & description		An Australian approach
	'catastrophic risk' and 'critical safety incident'); AI Act (EU), art 3 (definition of 'serious incident'); Law on Artificial Intelligence (Vietnam), art 3(8) (definition of 'serious incident'); Proposed AI Incident Reporting Act HR 9477 (US), s 2(b)(2) (definition of 'reportable activity'); Transport Safety Investigation Regulations 2021, regs 6 and 11A-12 (definitions of 'aircraft accident', 'aircraft incident' and 'reportable matters'); Work Health and Safety Act 2011, s 37 (defining 'dangerous incident'); National Health Security Act 2007, s 48 (definition of 'reportable events'); Therapeutic Goods Act 1989, ss 29A and 29AA (reportable information); Therapeutic Goods (Medical Devices) Regulations 2002, reg 5.7 (coverage of near misses); Australian Consumer Law, s 131 (definition of reportable incidents).	
2.4	Exemptions <i>Exemptions may be necessary to exclude legitimate use cases where reporting would be counterproductive, such as in academic research, defence and national security, with clearly defined limits.</i>	Any exemptions to Australia's AIIMR regime should be minimised and narrowly scoped. Two categories for potential exemptions are: (a) Non-commercial research by universities and public research bodies for the purposes of understanding existing deployed models, except where engaged by a frontier AI company as a third-party evaluator; (b) Models explicitly contracted for development and use by Australian government entities for national security and defence purposes. For national security and defence exemptions, there should be separate monitoring and reporting obligations at least as stringent as those that would apply under the AIIMR regime.
	Precedents: Responsible AI Safety and Education Act (New York), §1426 (academic research exemption); AI Act (EU), art 2(3), (6) (academic, national security and defence exemptions); AI Basic Act (Korea), art 4(2) (national security and defence exemption).	
3. Monitoring and reporting – how and when		
3.1	Duty to monitor and log <i>To ensure that reporting obligations do not incentivise entities to avoid monitoring for incidents, it is important to require covered entities to actively monitor and log their systems for potential reportable incidents.</i>	Covered entities should be required to document and maintain a system to monitor, detect, and record reportable incidents. The system should be proportionate to the entity's size and role, the relevant frontier AI model's capabilities, and how it is being used. This system should also assess suspected reportable incidents and record whether or not they were judged as reportable and why. The system should cover training, internal use, and deployment. It should log actions of AI agents, chain of thought and other reasoning transcripts, and reports from users, staff, and external researchers. This logging should include the ability to attribute actions to specific models and agent instances. It should actively monitor for harms and

	Element & description	An Australian approach
		potential incidents during deployment, including through interactions with other AI and computer systems. Logs and records created by AI agents also need to be protected from AI agents, given past attempts by AI agents to tamper with these records.
	Precedents: AI Act (EU), arts 55(1)(c), 72 (obligation to monitor and document); Code of Practice for General-Purpose AI Models (Safety and Security Chapter) (EU), Measure 9.2 (relevant information for tracking serious incidents); Transparency in Frontier Artificial Intelligence Act (California), §22757.12(a)(8) (obligation to publish framework including for identifying critical safety incidents); AI Basic Act (Korea), art 32(1) (obligation to have a system that monitors safety accidents); APRA CPS 234, para. 23 (obligation to have a security incident detection mechanism); APRA CPS 230, paras. 16, 32-33 (obligation to identify operational risk incidents and near misses). See also Z. Kara, Improving Frontier AI Incident Reporting Regimes (GovAI, Policy Brief, September 2026), pp. 10-12.	
3.2	Agency receiving reports <i>The agency receiving AI incident reports must have sufficient technical capabilities to understand and act upon reports. It must also have broad enough expertise to address the variety of potential threat models or refer reports to relevant regulators.</i>	Australia's AI Safety Institute (AISI) has the appropriate technical and governance capabilities to assess, refer, and act on AI incident reports. It is also connected to the international work of AISIs, enabling information and threat sharing. However, incorporating the AISI into an AIIMR will require providing it with greater statutory backing and resourcing. The Australian Signals Directorate's (ASD) Australian Cyber Security Centre (ACSC) is also a potential option for receiving reports, noting its existing role receiving cyber reports.³⁷ However, its integration within Australia's intelligence community may limit its ability or willingness to share reports more broadly. Its particular expertise in cyber security may also make it a less appropriate recipient of reports of broader AI model incidents (such as those relating to biological risk or harms to human health). Alternatively, a new standalone body could be made responsible for administering the AIIMR regime.
	Precedents: AI Act (EU), arts 55 and 73 (AI Office and market surveillance authorities); Transparency in Frontier Artificial Intelligence Act (California), §22757.13 (Office of Emergency Services); Responsible AI Safety and Education Act (New York), §1420(16) (an office within the Department of Financial Services).	

³⁷ Australian Signals Directorate, [Report a cyber security incident](#).

	Element & description	An Australian approach
3.3	Timing of reports <i>Consistent with similar regimes overseas and in other sectors, reporting deadlines should be tiered based on severity or likelihood of harm.</i>	Reporting timeframes should commence when a covered entity has facts sufficient to support a reasonable belief that a reportable incident has occurred. Specific timeframes may need to be co-designed with industry. However, an initial report could be required within: (a) Twenty-four hours, where there is an imminent risk of, or actual, death or serious injury; (b) Forty-eight hours, where there is a widespread or systemic risk, including of critical infrastructure disruption; (c) Seventy-two hours, for all other reportable incidents. Incidents could also be made reportable immediately or as soon as reasonably practicable where harm is imminent or has occurred. Comprehensive reports could then also be required to follow within 15 days, providing entities sufficient time to collect and analyse all relevant information. Final reports, lodged within 30 days of the incident being resolved, should include causal analyses and remediation measures undertaken.
	Precedents: AI Act (EU), art 73 (initial report within 15 days, 2 days for widespread infringement, 10 days for death); Transparency in Frontier Artificial Intelligence Act (California), §22757.13(c) (initial report within 15 days or 24 hours to authorities if imminent risk of death or serious injury); Responsible AI Safety and Education Act (New York), §1422(3) (initial report in 72 hours or 24 hours to authorities if imminent risk of death or serious injury); Security of Critical Infrastructure Act 2018, ss 30BC-30BD (cyber incident reports within 12 or 72 hours depending on severity).	
3.4	Contents of reports <i>Standard form reporting will help ensure reports are comparable, able to be easily analysed and shared, and that industry is clear on what is required. These should also be interoperable with overseas regimes (to the extent possible).</i>	The precise contents of reports should be co-designed with industry. The contents will also vary between initial, comprehensive and final reports. Common categories of reportable information for incidents, including under OECD guidance, have included: (a) Metadata: title, description, date, countries concerned, submitter details; (b) Technical information: model name and

	Element & description	An Australian approach
		version, agent identifiers, developer information, extent of deployment; (c) Harm: type, scale, involvement of the AI model, and why the incident is reportable; (d) Affected people, entities or systems; and (e) Remediation undertaken or planned. Access to system logs, transcripts, or the model itself may also be required in some circumstances. These should be preserved and access enabled upon request. The occurrence of an incident should also trigger preservation obligations to suspend ordinary deletion or overwriting procedures of relevant records.
	Precedents: Incident Report for Serious Incidents under the AI Act (High-risk AI systems) (Draft Reporting Template) (EU); Report for Serious Incidents under the AI Act (General-Purpose AI Models with Systemic Risk) (Draft Reporting Template) (EU). See also Z. Kara, Improving Frontier AI Incident Reporting Regimes (GovAI, Policy Brief, September 2026), 13; and OECD, Towards a common reporting framework for AI incidents (2025).	
3.5	Regulatory overlap <i>An AIIMR will overlap with other regulatory regimes, such as for critical infrastructure operators, or in relation to cyber security incidents. Duplicate reporting should be minimised to avoid unnecessary costs and inconsistent responses.</i>	Where reporting entities are covered by equivalent reporting requirements, compliance with those requirements should be deemed compliance with the AIIMR. To provide clarity and certainty, the equivalent regimes should be specified in rules or a Ministerial determination. This should be supported by information-sharing between regulators to ensure there is a central source compiling and tracking AI incidents in particular.
	Precedents: AI Act (EU), arts 73(9)-(10); Transparency in Frontier Artificial Intelligence Act (California), §22757.13(h)-(i); Responsible AI Safety and Education Act (New York), §1422(9); Australian Consumer Law, s 131(2)(c).	
4. Response — how reports are used		
4.1	Use of reports <i>There should be clear guidance on what incident reports will be used for, in line with the purpose of the AIIMR regime.</i>	AI incident reports should be primarily used for three purposes: (a) Referral to response agencies to activate emergency and regulatory responses; (b) Review of significant incidents to recommend mitigations to avoid recurrence or escalation of harms in future, including through investigation or ordered investigation of incidents; and

	Element & description	An Australian approach
		(c) Remediation of incidents, including through the use of specific powers to direct remediation and prevention action by the reporting entity. Reporting should not grant safe harbour, and other regulators should be able use the content of shared reports to implement other statutes.
	Precedents: AI Act (EU), arts 73(7)-(8) (referral of reports to authorities and intervention where necessary); Transparency in Frontier Artificial Intelligence Act (California), §22757.13(e), (g) (referral of reports to agencies and public reporting); Transport Safety Investigation Act 2003, ss 25, 25A (publication of reports on incidents and recommendation of safety actions).	
4.2	Response powers <i>To enable effective remediation and response, new powers will be needed for the agency responsible for handling AI incident reports.</i>	The agency receiving AI incident reports will need powers to: (a) Compel information, documents, testimony, and access to models and relevant model records (such as chain-of-thought transcripts or evaluation logs); (b) Conduct or order independent investigations of sufficiently serious incidents (including near misses); (c) Issue safety bulletins and recommendations to industry and users; and (d) Require corrective action plans or direct covered entities to restrict, modify or withdraw a model, or take other action to protect the public or respond to an incident.
	Precedents: AI Act (EU), arts 91-93 (powers to request information, evaluate AI models, request AI companies take certain measures); AI Basic Act (Korea), art 40 (power to order data production, conduct investigations, or order corrective actions); Law on Artificial Intelligence (Vietnam), art 12(3) (power to order suspension or recall); Transport Safety Investigation Act 2003, Part 5 (investigation powers); Security of Critical Infrastructure Act 2018, Part 3A (intervention powers); Australian Consumer Law, ss 122, 129 (recall and safety warning notice powers).	
4.3	Information sharing <i>AI incidents will often cut across sectors and regulators. Information sharing between regulators, and with partners overseas, will be needed to ensure an accurate picture of the threat landscape can be constructed.</i>	The Australian AIIMR agency should be able to: (a) Share relevant reports with other agencies and regulators, with appropriate protections for commercial-in-confidence data or information; (b) Share reports and analysis with overseas regulators and AISIs, subject to confidentiality arrangements where necessary; (c) Produce public reporting of de-identified

Element & description		An Australian approach
		incidents and lessons learned; and (d) Provide annual reports to Parliament on incidents, responses, and trends.
Precedents: AI Act (EU), art 78(5) (ability to exchange information with other countries under confidentiality arrangements); Transparency in Frontier Artificial Intelligence Act (California), §22757.13(g) (annual reports); Therapeutic Goods Administration's Database of Adverse Event Notifications ; Office of the Australian Information Commissioner's Notifiable Data Breaches Publications .		
5. Protections and enforcement		
5.1	Confidentiality and limits on use of reports <i>Reporting entities should be incentivised to report fully, honestly and swiftly. This requires protections for confidential information and against self-incrimination, while not preventing legitimate regulatory action against entities.</i>	Australia's AIIMR regime should consider drawing on precedents in similar regimes, such as: (a) Exemptions from freedom of information and other disclosure requirements; (b) Protections that maintain legal professional privilege over disclosed materials; and (c) Clarification that AI incident reports are not an admission of fault or liability.
Precedents: AI Act (EU), art 78(5) (confidentiality protections); Transparency in Frontier Artificial Intelligence Act (California), §22757.13(f) (public records exemption); Cyber Security Act 2024, s 31 (protections for legal professional privilege).		
5.2	Whistleblower protections <i>Protections for whistleblowers can help incentivise reporting entities to develop a culture of reporting by preventing retaliation against staff reporting failures or harms.</i>	Australia has existing whistleblower protections, particularly in relation to misconduct. These should be reviewed to determine if amendments are needed to ensure they provide sufficient protections for reports regarding AI incidents, such as: (a) Protection from retaliation from the reporting entity; (b) Requiring reporting entities to have internal reporting channels; (c) Voiding contractual terms that prevent disclosures; and (d) Potential protection from prosecution for certain categories of reports.
Precedents: AI Act (EU), art 87 (application of EU whistleblowing directive to the AI Act); Labor Code (California), §1107.1 (whistleblower protections for frontier AI company employees); Corporations Act 2001, Part 9.4AAA (corporate whistleblower protections); Public Interest Disclosure Act 2013 (public service whistleblower protections).		
5.3	Penalties and enforcement <i>Penalties for non-compliance with the reporting regime should be sufficient to</i>	Civil penalties should apply for failing to monitor or report covered incidents adequately, and for missing timeframes for reporting. Increased civil

	Element & description	An Australian approach
	<i>ensure they are not seen as just a cost of doing business.</i>	and criminal penalties should apply for providing false or misleading information, for repeated breaches, or gross or wilful negligence resulting in non-compliance. Precise penalties will depend upon the nature and circumstance of any breach. However, maximum applicable penalties should be set at a level to ensure obligations carry real weight. For example, under Australia’s therapeutic goods regulatory regime, penalties of up to 30,000 penalty units for body corporates and twelve months’ imprisonment can apply for failing to notify adverse events within specified timeframes.³⁸
	Precedents: AI Act (EU), art 101 (fines of up to 3% of annual worldwide turnover of €15 million for general-purpose AI providers); Transparency in Frontier Artificial Intelligence Act (California), §22757.15 (civil penalties of up to US\$1 million per violation); Responsible AI Safety and Education Act (New York), §1427 (fines of up to US\$1 million for first violations, and US\$3 million for subsequent violations).	
6. Implementation		
6.1	New or amending legislation <i>Elements of an AIIMR regime could be implemented through amendments to existing regulatory regimes (such as in relation to data centres under the Security of Critical Infrastructure Act). However, to ensure a cohesive and consistent approach is taken, it is likely preferable for a separate, standalone regime to be established.</i>	An Australian AIIMR could be instituted via the legislated Australian Standards for AI, if these covered both training and inference, or through separate legislation. Obligations on data centres could be implemented through amendments to the Security of Critical Infrastructure Act, although this risks fragmenting reporting obligations under both regimes. Any legislation should establish the framework for the reporting regime but leave specific thresholds and definitions to delegated instruments. This will ensure they can be regularly reviewed to keep pace with technological developments.
	Precedents: Biosecurity Act 2015, s 154 (definition of ‘reportable biosecurity incident’ can be set through legislative instrument); Cyber Security (Ransomware Payment Reporting) Rules 2025, rule 6 (sets turnover threshold for reporting entities); Security of Critical Infrastructure Act 2018 (covered assets and other elements set by rules and other instruments).	

³⁸ Therapeutic Goods Act 1989 (Cth), ss 29A, 29AA.