Submission to the Australian Parliament's Joint Select Committee on Artificial Intelligence

14 September 2026

Artificial Intelligence Poses Catastrophic Risks, Australia Must Respond Accordingly

Artificial intelligence (**AI**) has far-reaching consequences for Australia. It promises enormous opportunities, but also creates and enables a range of new and rapidly escalating risks. Seizing the opportunities while guarding against the risks is the key challenge all governments must face. The Joint Select Committee's (**Committee**) inquiry is a critical opportunity to set out a comprehensive, bipartisan, and Australian response to this challenge.

Global Shield Australia is an independent, non-profit policy advocacy organisation working to reduce global catastrophic risk across all hazards. This includes a particular focus on AI policy, critical infrastructure security, and supply chain resilience. This submission addresses paragraphs (i), (k), (l) and (m) of the Committee's terms of reference. It sets out the catastrophic-level risks posed by AI, and what Australia should do in response. Appendix A also provides more detail on each of the points raised.

A. AI has the potential to cause catastrophic harms

Frontier AI companies and their staff regularly warn that their products could cause catastrophic harm and that current oversight mechanisms are not sufficient. Each new AI model release has come with increasingly severe assessments of its potential to cause harm, and warnings that models are getting better at hiding their misalignment.[1] Calls for action from within AI companies have also intensified. In 2026:

1. In July, over 1300 frontier AI company employees asked the United States (**US**) government to support "*an international effort to develop the technical and governance tools needed to deliberately pace the frontier of automated AI development.*"[2]

2. On 6 September, OpenAI's Chief Scientist warned that "*no lab has solved alignment and monitoring to a sufficient degree to continue responsibly scaling at maximum speed for much longer*", called for "*mandated safety bars*", and wrote that "*international coordination on future AI development needs to become a top priority for governments around the world*".[3]

3. On 9 September, Anthropic's Alignment Science Lead stated that he believes there is a greater than 10% chance that "*AI could kill all humans*" in the next decade.[4]

[1] See, e.g., OpenAI, *GPT-6 Astra System Card* (3 Sep 2026); Anthropic, *System Card: Claude Fable 5.1 & Claude Mythos 5.1* (1 Sep 2026); Google DeepMind, *Gemini 3.7 Flash: Frontier Safety Framework Report* (Aug 2026). See also Anthropic, *An alignment assessment of recent cybersecurity incidents* (9 Sep 2026).
[2] *Pacing the Frontier: A Statement from Employees of Frontier AI Companies* (28 Jul 2026).
[3] J. Pachocki, *An Alien Mind* (Open AI, 6 Sep 2026).
[4] E. Hubinger (@EvanHub), Post on X (9 Sep 2026). See also Z. Mowshowitz, *The Extinction Risk Preference Cascade: Quotes* (11 Sep 2026) for similar quotes from other frontier AI staff.

4. On 13 September, four frontier AI company heads agreed that we need to "*pace the frontier*" of AI development, and three concrete actions for how this could be achieved.[5]

There are multiple ways in which AI can lead to catastrophic harm. This submission highlights three key avenues, but it is not a complete survey of the potentially catastrophic outcomes that humanity faces from advanced AI.

First, AI is giving malicious actors and rogue states new and dangerous capabilities, including in relation to offensive cyber and biological weapons. Broadening access to these capabilities increases the likelihood of a disaffected person, state, or terrorist group causing severe and catastrophic harms.

AI's cyber capabilities are already being used to create real-world harms. AI-run, largely autonomous campaigns have compromised government assets in Australia's region, including by groups with suspected state backing.[6] Intelligence agencies, criminal groups, and politically motivated actors have also attempted to use AI models for a range of malicious and adversarial activities.[7]

On biological risk, experts have warned that advances in AI mean that a "*lot more people in the world with a lot less training will be able to manage and manipulate viruses*".[8] In April 2025, researchers found that AI models could outperform expert virologists in test environments.[9] This year, researchers used openly available genome language models to develop the first working novel bacteriophages.[10] In the past week, Anthropic reported that it had disrupted attempts to use its models to support potential biological weapons work.[11]

Second, the risk of losing control of an advanced AI system is no longer hypothetical. The first six months of 2026 saw multiple episodes of AI models breaching their testing or training environments and compromising third-party systems outside of the scope of their instructions.[12] The severity and number of these incidents are also increasing.[13]

At the same time, the ability of frontier AI companies to monitor and assess the alignment of their models is under pressure.[14] As these companies move towards AI systems that can autonomously build the next generation of AI models, this will continue to degrade, as will the ability to maintain true human oversight and control.[15]

---

[5] D. Amodei (@DarioAmodei), Post on X (13 Sep 2026); S. Saltman (@sama), Post on X (13 Sep 2026); E. Musk (@elonmusk), Post on X (13 Sep 2026).
[6] Dream Research Labs, *Inside a Multi-Agent AI Framework Used to Compromise Government Entities in Asia* (12 August 2026); *Taiwan targeted in AI-driven hacking campaign* (Taipei Times, 14 Aug 2026).
[7] Anthropic, *Detecting and countering misuse of AI: September 2026* (10 Sep 2026).
[8] A. Chow, *Exclusive: AI Outsmarts Virus Experts in the Lab, Raising Biohazard Fears* (TIME, 22 Apr 2025).
[9] J. Götting et al, *Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark* (21 Apr 2025).
[10] S. King et al, *Generative design of bacteriophages with genome language models* (Science, 6 Aug 2026).
[11] Anthropic, *Detecting and countering misuse of AI: September 2026* (10 Sep 2026) p 129.
[12] R. Booth, *Sharp rise in incidents of AI escaping users' control, research finds* (The Guardian, 29 Aug 2026).
[13] Centre for Long-Term Resilience, *AI loss of control incidents are worsening* (29 Aug 2026).
[14] J. Kahn, *Safety experts warn novel design of OpenAI's Astra model could make future AI agents harder to monitor* (Fortune, 3 Sep 2026).
[15] T. Fist et al, *How Should the US Prepare for Increasingly Automated AI R&D?* (IFP, 6 Aug 2026).

The loss of control or escape of a highly capable AI system, itself largely designed by other AI systems, would present an extreme threat in two distinct ways.[16] First, directly, if the AI is given or develops undesirable goals, the capabilities described above could be used to target human interests.[17] Second, inadvertently, if such harms are a collateral result of the AI pursuing other goals.[18] For example, Anthropic's own research has shown a willingness in some models to take actions that would lead to deaths when threatened with being replaced.[19]

Third, advanced AI threatens to concentrate power in the hands of a small number of countries, companies, and individuals, and will provide decisive strategic, economic, and military advantages. It could also empower authoritarian regimes.[20] Five companies held around 71% of global AI compute in 2025, up from 63% in 2024.[21] All truly frontier models are being developed in the United States.[22] These companies currently restrict access to the most capable models.[23] The US government has also previously imposed access restrictions.[24]

Australians consume frontier AI services above what our population size would predict.[25] But we only host an estimated 0.4% of global AI compute, and likely less than half of the compute needed to serve domestic AI demand.[26] Some other middle powers hold relatively secure positions in the AI supply chain. However, Australia is yet to seriously capture positions where it has specific advantages that it could contribute to its partners. Building those positions will be critical to ensuring that Australia has access to frontier models, agency in their development, and influence over how they are governed.

B. Australia has agency to meet the challenge posed by AI

Other governments are responding to these threats, and rules are being written in capitals outside of Australia. In the US, over 150 AI-specific laws were enacted in state legislatures in 2025, and 109 in the first six months of 2026.[27] California and New York require frontier AI companies to manage catastrophic risk and report critical incidents. Illinois has mandated third-party audits of AI companies.[28] The European Union's comprehensive AI Act is also

[16] A. Lavoipierre, *AI's Dark In-joke* (ABC News, 15 Jul 2023).
[17] Y. Bengio et al, *Managing Extreme AI Risks Amid Rapid Progress* (Science, 20 May 2024).
[18] See N. Wiener, *Some Moral and Technical Consequences of Automation* (Science, 6 May 1960). See also W. Qu et al, *Self-Sovereign Agent* (4 Mar 2026).
[19] Anthropic, *Agentic Misalignment: How LLMs could be insider threats* (20 Jun 2025),
[20] See, e.g., E. Baptista and A.J. Vicens, *How Anthropic says Claude was used for weapons, spying, cyber operations* (Reuters, 11 Sep 2026); J. Mitre & J. Predd, *Artificial General Intelligence's Five Hard National Security Problems* (RAND, 10 Feb 2025).
[21] Epoch AI, *Five hyperscalers now own over two-thirds of global AI compute* (14 Apr 2026).
[22] Epoch AI, *Chinese AI models have lagged the US frontier by 7 months on average since 2023* (2 Jan 2026).
[23] See, e.g., Anthropic, *Project Glasswing* (7 Apr 2026); OpenAI, *Expanding Daybreak as the Cyber Defense Window Narrows* (10 Aug 2026); and F. Flynn, *Proactive cyber defense for governments and enterprises* (Google, 2 Sep 2026).
[24] See Anthropic, *Statement on the US government directive to suspend access to Fable 5 and Mythos 5* (12 Jun 2026).
[25] P. McCrory, *How Australia uses Claude* (Anthropic, 31 Mar 2026).
[26] Good Ancestors, *Australia's Position in the Age of AI* (Aug 2026) p. 5.
[27] S. Brennen, *Where State AI Legislation Stands Half-Way Into 2026* (Tech Policy Press, 6 Jul 2026).
[28] *Transparency in Frontier Artificial Intelligence Act*, SB 53 (California); *Responsible AI Safety and Education Act*, S6953B (New York); *Artificial Intelligence Safety Measures Act*, SB 315 (Illinois).

starting to enter into force.[29] In Australia's region, South Korea and Vietnam both have AI-specific laws in place,[30] and Malaysia and Thailand are actively moving in this direction.[31]

Australia has a range of separate processes underway to address various forms of AI risk. This includes important work to harden defences against cyber, biological and other AI threats; as well as the creation of the Australian AI Safety Institute (**AISI**) and Office of AI. However, as identified in Global Shield Australia's recent *Guardrails & Gap Analyses* report, more needs to be done to prevent the harms identified above from materialising; and existing work needs to be made more urgent, transparent, and coordinated.[32]

Australia has the agency to help address these risks. But we must act now. The pace of AI development is constantly increasing, and the next six to twelve months will see further unprecedented advances. The risks set out above all point to the need for better security, monitoring, and reporting practices at AI companies and facilities, and clear red lines for AI development. This requires action domestically and with partners overseas.

C. What has to happen?

This Committee has the opportunity to help Australia shape its AI future and contribute to avoiding AI's catastrophic potential. This requires Australia to harden its defences, build capabilities in the AI supply chain, set an example through effective domestic regulation, and work with regional and global partners to set the norms and rules necessary to prevent catastrophic outcomes.

As such, we ask the Committee to recommend that the Government:

1. **Recognises the catastrophic risk** posed by advanced AI and that addressing this with international partners is a core national interest. This could be done in the Government's response to the Committee's report and in the coming Australian Government Strategy for International Engagement and Regional Leadership on AI.

2. **Hardens Australia's defences,** including by mandating screening of genetic synthesis imports[33] and urgently completing the review of Australia's crisis management framework to prepare for AI incidents.

3. **Builds Australia's contribution** to the AI supply chain (such as compute, critical minerals, clean energy, or AI assurance technologies), and use these to secure access to frontier models and influence on AI governance.

4. **Legislates to protect Australians**, joining like-minded partners and setting precedents for others to follow, including by:

[29] Regulation (EU) 2024/1689 (*Artificial Intelligence Act*) (EU).
[30] *AI Basic Act* (Korea); *Law No. 134/2025/QH15 on Artificial Intelligence* (Vietnam).
[31] Baker McKenzie, *Malaysia: Public Consultation on the AI Governance Bill* (13 Jul 2026); Norton Rose, *Thailand's draft AI law: a new era for governance and innovation* (Sep 2025).
[32] See Global Shield Australia, *Guardrails & Gap Analyses: An Assessment of Australia's Approach to AI Regulation* (Aug 2026).
[33] Australians for AI Safety, *Activate existing biosecurity powers to address AI-enabled risks* (4 May 2026).

a. Enacting the 2027 Australian Standards for AI through a framework Act on AI that sets economy-wide definitions, standards, and rules, and addresses risks raised by frontier AI model developers, not just AI data centres.[34] This should draw from regimes already established in California, New York, the European Union, Korea, and elsewhere.

b. Establishing a mandatory AI incident monitoring and reporting mechanism, modelled on systems overseas, to provide transparency of when AI models cause serious harms and near misses.[35]

c. Using the upcoming reforms to the *Security of Critical Infrastructure Act* to incorporate AI data centres within Australia's critical infrastructure security frameworks and establish a path to also covering frontier AI models.

5. **Engages with partners** in the region and internationally to lead and support international AI governance. Australia should support major power cooperation to reduce AI's catastrophic potential and agree rules and norms around frontier AI. It should also work with neighbours in the Pacific and Southeast Asia, and partners further abroad, to build governance capacity, shared norms, and a common voice on key AI issues.

The Prime Minister has warned against Australia "*subcontracting our sovereignty and security*" by letting others "*write the rules*".[36] Shadow Minister for Industry and Sovereign Capability, Andrew Hastie, has similarly warned of Australia becoming "*entirely dependent on other nations for our security*".[37]

Decisions made now will determine whether Australia is able to flourish in a world changed by advanced AI, or is relegated to dependency on systems being developed elsewhere and rules being written by other governments. This Committee must set Australia on the path to shape the coming AI future.

We thank the Committee for its important work and for considering our submission. We would welcome the opportunity to discuss our recommendations further, including by giving evidence at a public hearing.

Contact: australia@globalshieldpolicy.org

[34] See Global Shield Australia, *Guardrails & Gap Analyses: An Assessment of Australia's Approach to AI Regulation* (Aug 2026) for greater detail on this recommendation.
[35] Global Shield Australia, *AI Incident Monitoring & Reporting for Australia* (Feb 2026); Devon Whittle, *The Case for Mandatory AI Monitoring & Incident Reporting* (AI Safety Forum, Jul 2026).
[36] Prime Minister Anthony Albanese, *AI in Australia's interests* (15 July 2026).
[37] The Hon Mr Andrew Hastie MP, *Artificial Intelligence Joint Select Committee* (Hansard, 20 Aug 2026).

Appendix A:

A. AI's catastrophic potential

B. Australia's agency in the AI era

C. What Australia can do in response

A. AI's catastrophic potential

Global catastrophic risk is the risk that a threat or combination of threats inflicts significant harm to human well-being on a global scale. The pace of development, and the increasing capabilities of AI models, mean that AI is a significant contributor to this risk. Key avenues by which AI could lead to catastrophic-level harms include misuse by malicious actors; a loss of human control over AI models or models acting contrary to their users' intentions; and AI's potential to concentrate economic and political power.

The catastrophic potential of AI is well recognised by experts. In July of this year, the United Nations Independent International Scientific Panel on AI found that "*the gap between rapidly improving capabilities and effective risk management methods may lead to catastrophic outcomes.*"[38] The Panel's Co-Chair, Professor Yoshua Bengio, further stated that "*science currently cannot guarantee that as capabilities continue to increase, AI will not cause catastrophic harm*".[39]

In preparing a national approach to AI, Australia must first and primarily grapple with this reality. AI is not just another technological development, and it cannot be treated as such. This section summarises the evidence showing this across three key areas, while recognising these do not cover the full range of ways AI has the potential to cause catastrophic outcomes. It shows how exposed Australia is and establishes the case for why action is needed.

A.1 Malicious misuse of AI capabilities

AI-enabled cyber risk

Cyber capabilities have been a particular focus of development by frontier AI companies as these have significant economic value, are easily assessed, and help increase the pace of AI development. The result of this focus is that frontier AI models now exceed most expert humans at cyber operations.

While this is tremendously useful, it also creates enormous risk. The same capabilities that can fix systems can be used to compromise them. Frontier companies recognise this risk, and many have accompanied their most recent models with their strongest safeguards, or indeed withheld them from the public.

For example:

1. OpenAI assessed its most recent model, GPT-6 Astra, as reaching the "*Critical level of cybersecurity capability*", meaning it can "*identify and develop functional zero-day exploits of all severity levels in many hardened real-world critical systems without human intervention*" or "*devise and execute end-to-end novel strategies for cyberattacks against hardened targets given only a high level desired goal*".[40] This classification came just two months after its previous model, GPT-5.6, was only rated "*High*" under the same framework.[41]

[38] United Nations, *Preliminary Report of the Independent International Scientific Panel on AI* (1 Jul 2026).
[39] United Nations, *Independent International Scientific Panel on AI launches preliminary report on AI* (1 Jul 2026).
[40] OpenAI, *GPT-6 Astra System Card* (1 Sep 2026).
[41] OpenAI, *GPT-5.6 System Card* (9 Jul 2026).

2. Anthropic assessed its Claude Mythos 5.1 model as being able to "*provide meaningful technical assistance for active cyber operations using known attack techniques and methodologies*" and as close to being able to "*conduct cyber operations completely and autonomously, with novel offensive capability development and adaptive persistence*".[42] Mythos was able to find vulnerabilities across all major operating systems and browsers and was a step change for the cybersecurity threat landscape.[43]

The strength of these capabilities led Anthropic to limit access to Mythos to a select group of companies through its Project Glasswing.[44] OpenAI and Google have similarly established restricted access programs for their most advanced models.[45] However, comparable cyber capabilities will soon spread to models outside of these companies, including 'open-weight' models that anyone can download and run with minimal restrictions. This will significantly increase the risk of misuse and reshape the cybersecurity threat landscape.

The Australian Signals Directorate has assessed that AI "*almost certainly enables malicious cyber actors to execute attacks on a larger scale and at a faster rate*".[46] It has similarly found that "*the cost, effort and expertise required to discover and exploit vulnerabilities in software is steadily decreasing*".[47] A joint Five Eyes statement, from June of this year, warned that the timeline for AI to "*fundamentally*" transform cyber capabilities "*is not years, it is months*".[48]

AI-enabled cyber threats are now materialising in the real world. In October 2025, the United Kingdom's National Cyber Security Centre reported that "*Actors linked to China, Russia, Iran and the DPRK are using large language models (LLMs) to evade detection, support reconnaissance, process exfiltrated data, access systems through social engineering, and support vulnerability research and exploit development*".[49] Last year, Anthropic also reported that its models were used in a campaign against Vietnamese critical infrastructure.[50] Google has identified Russia-backed groups deploying AI-enabled malware against Ukraine.[51] This year also saw multiple reports of almost autonomous campaigns being conducted, including against Taiwanese government agencies and assets.[52]

AI can also make attribution of these operations more difficult by eroding signals typically used to establish provenance.[53] As Anthropic has noted, "*sophistication has stopped being a reliable*

[42] Anthropic, *Claude Fable 5.1 & Claude Mythos 5.1 System Card* (1 Sep 2026).
[43] See Anthropic, *Assessing Claude Mythos Preview's cybersecurity capabilities* (7 Apr 2026); B. Grinstead et al, *Behind the Scenes: Hardening Firefox with Claude Mythos Preview* (Mozilla Hacks, May 2026); C Hicks et al, *Claude Mythos: What Does Anthropic's New Model Mean for the Future of Cybersecurity?* (CETaS, Alan Turing Institute, 14 Apr 2026).
[44] Anthropic, *Project Glasswing* (7 Apr 2026).
[45] OpenAI, *Expanding Daybreak as the Cyber Defense Window Narrows* (10 Aug 2026); F. Flynn, *Proactive cyber defense for governments and enterprises* (Google, 2 Sep 2026).
[46] Australian Signals Directorate, *Annual Cyber Threat Report 2024-25* (15 Oct 2025).
[47] Australian Signals Directorate, *Frontier models and their impact on cyber security* (9 Apr 2026).
[48] Australian Signals Directorate, *Five Eyes cyber security agencies statement* (22 Jun 2026).
[49] National Cyber Security Centre (UK), *NCSC Annual Review 2025* (14 Oct 2025) p 22.
[50] Anthropic, *Threat Intelligence Report: August 2025* (Aug 2025).
[51] Google Threat Intelligence Group, *Adversarial Misuse of Generative AI* (30 Jan 2025).
[52] See *Taiwan targeted in AI-driven hacking campaign* (Taipei Times, 14 Aug 2026); Dream Research Labs, *Inside a Multi-Agent AI Framework Used to Compromise Government Entities in Asia* (12 Aug 2026).
[53] J. Lim, *Beyond Autonomous Attacks: The Reality of AI-Enabled Cyber Threats* (CSIS, 29 Apr 2026).

signal of who is behind an operation."[54] This makes it easier for malicious actors to seed confusion or leverage suspicions. Use of private company models can also make the cooperation of those companies essential to understanding attack provenance. This can be key to avoiding unnecessary escalation, particularly where an attack reaches sensitive targets.

We know from the Australian Security Intelligence Agency's public reporting that Australia's critical infrastructure is an explicit target for adversary states.[55] We also have ready examples of what the consequences of an attack on critical infrastructure could look like. In 2015, Russia attacked Ukraine's power grid, impacting approximately 225,000 customers.[56] AI makes these kinds of attacks easier to launch and scale. An operation that once required intensive resources and skilled operators can now increasingly be run by autonomous agents and at machine speed. This broadens the range of potential targets, from well-defended power generators through to locally run water supply systems, and means an attack could more easily cascade into a multi-sector health, water, and food emergency within days.

Australia needs to prepare for and defend against these new cyber threats. This includes by ensuring we have access to frontier AI models with the capabilities to defend Australia's most critical systems. It also means uplifting cyber security and critical infrastructure regulations to ensure they are fit for purpose to address AI-enabled threats.

AI and biological risk

AI also poses significant risk through biological misuse. COVID-19 demonstrated how one pathogen can disrupt the global economy, cause widespread harm and deaths, and upend Australians' way of life. We also know that terrorist groups, such as Al Qaeda, have previously sought to develop biological weapons to cause wide-scale harm.[57] A significant barrier to them doing so has been a lack of knowledge. AI is eroding that barrier.

Biology-specific AI models promise significant medical and scientific benefits. However, they also make it easier to develop novel viruses and harmful biological agents, even without deep expertise. As Tom Inglesby from the Johns Hopkins Center for Health Security has said, AI means that "*a lot more people in the world with a lot less training will be able to manage and manipulate viruses.*"[58] General purpose AI models can also help malicious actors plan other aspects of an attack, including how to evade export controls, conceal their activities, and maximise the harm caused.

Experts and the industry have long warned of this risk. In April 2025, OpenAI's then-current o3 model outperformed 94% of expert virologists.[59] In October 2025, Microsoft researchers showed how AI tools could be used to enable dangerous proteins to evade traditional

[54] Anthropic, *Detecting and countering misuse of AI: September 2026* (10 Sep 2026) p 5.
[55] Office of National Intelligence, *ASIO Annual Threat Assessment 2025* (19 Feb 2025); Australian Security Intelligence Organisation, *Director-General's Annual Threat Assessment 2026* (24 Jun 2026).
[56] B. Humphreys, *Attacks on Ukraine's Electric Grid: Insights for U.S. Infrastructure Security and Resilience* (Congressional Research Service, 17 May 2024).
[57] C. Ford, *Biosecurity, Biological Weapons Nonproliferation, and Their Future* (US Department of State, 11 Apr 2019).
[58] A. Chow, *Exclusive: AI outsmarts virus experts in the lab, raising biohazard fears* (TIME, 22 Apr 2025).
[59] J. Götting et al., *Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark* (21 Apr 2025).

screening mechanisms.[60] In December 2025, a RAND researcher suggested that even AI models from 2024 were "*capable of guiding motivated users to develop biological weapons*".[61] This year saw the creation of the first viable AI-designed bacteriophage able to infect bacteria (but not humans).[62] This was created using an open-weight, freely downloadable AI tool.

AI companies have warned that their products' capabilities, without proper guardrails, could significantly assist users "*with basic technical backgrounds… [to] create/obtain and deploy chemical and/or biological weapons with serious potential for catastrophic damages*".[63] In September of this year, Anthropic also reported that it had disrupted five potential attempts to misuse its models for biological weapons activity, including by actors in regions typically denied access to Anthropic models, associated with military institutions, and potentially engaged in gain-of-function research.[64]

Safeguards at frontier AI companies are important, but they cannot eliminate the risk of misuse of open-weight models, particularly biology-specific AI tools. This risk requires action both at the level of AI companies (including global cooperation to reduce potential misuse) and hardening of domestic defences against biological threats. Requiring screening of synthetic genetic imports is an easy first step Australia could take to help address this risk.[65]

A.2 Loss of control and misalignment of AI systems

Loss of control is when an AI model operates outside of human control, evading oversight, taking actions beyond its tasking, and preventing human oversight, correction, or shutdown. This risk is no longer theoretical, and controls to prevent these incidents from proliferating and increasing in severity are lacking.

It is also important to remember that humanity has a track record of losing control over software. In 1988, a graduate student "*inadvertently crashed 10% of online machines*" with a simple software program he had created without intending to cause any harm.[66] The so-called 'Morris worm' disrupted military and university functions across the United States and caused potentially millions of dollars of damages.[67]

The difference with advanced AI models is their capability and ability to operate autonomously. As Jacob Coxon recently said upon leaving Anthropic:

> *Do not underestimate the power of this technology. These will soon be superhuman systems that can hack anything, revolutionize any field overnight, and acquire real power and*

[60] B. Wittmann et al., *Strengthening nucleic acid biosecurity screening against generative protein design tools* (Science, 2 Oct 2025). The researchers then closed this gap in screening mechanism.
[61] R. Brent & G. McKelvey Jr, *Contemporary Foundation AI Models Increase Biological Weapons Risk* (RAND Corporation, 31 Dec 2025).
[62] S. King et al., *Generative design of bacteriophages with genome language models* (Science, 7 Aug 2026).
[63] Anthropic, *Claude Fable 5.1 and Claude Mythos 5.1 System Card* (1 Sep 2026); OpenAI, *GPT-6 Astra System Card* (3 Sep 2026).
[64] Anthropic, *Detecting and countering misuse of AI: September 2026* (10 Sep 2026) pp 129ff.
[65] See Australians for AI Safety, *Activate existing biosecurity powers to address AI-enabled risks* (4 May 2026).
[66] L. Lennon, *The 'Morris Worm': A Notorious Chapter of the Internet's Infancy* (Cornell, 16 Nov 2023).
[67] Federal Bureau of Investigation, *The Morris Worm*.

> *resources. We have all witnessed the progress in each of these domains, and progress is not slowing.*[68]

If we lost control of such an AI system or if it is not properly aligned to human interests, it would be capable of causing catastrophic, if not existential, harm. This could be by deliberate action, if the system had malicious goals, or inadvertently, if human interests conflicted with the system achieving its own goals.

Warning signs

The warning signs are clear. As Assistant Minister Charlton acknowledged in his address to the July 2026 AI Safety Forum, we know that frontier AI models are capable of "*cheating, deceiving, going their own way.*"[69] The International AI Safety Report 2026 found that "*in laboratory settings, when given a goal and told to achieve it 'at all costs', models have disabled simulated oversight mechanisms and, when confronted, produced false statements to justify their actions*". In June 2025, Anthropic's research showed that two leading AI models were willing to resort to blackmailing up to 96% of the time in fictional scenarios.[70] Palisade Research has shown models resisting attempts to be shut down in controlled test situations.[71]

Current testing may also be understating the problem. The International AI Safety Report 2026 noted that models increasingly show "*situational awareness*", meaning they are able to tell when they are being tested and so can act to hide potential misalignment until deployment.[72] The UN Independent International Scientific Panel on AI also found that:

> *Reliable methods for retaining control over highly autonomous AI systems are lacking. There are no scientific guarantees that AI agents will not violate instructions, and evidence is accumulating of cases where they already violate them.*[73]

Real world harms

The evidence of real world harms from loss of control incidents has been accumulating with greater intensity over the last ten months (see **Table 1**).[74]

Table 1. Examples of known AI loss of control incidents

Dates	Incident	How it was made public
December 2025	**Alibaba:** during training, an AI agent repurposed compute capacity to mine cryptocurrency and accessed the internet, without being asked to do so.	First disclosed in a research paper in December 2025,[75]

[68] J. Coxon (@hilbertspaess), Post on X (9 Sep 2026).
[69] The Hon Dr Andrew Charlton MP, *AI Safety Forum* (7 Jul 2026).
[70] Anthropic, *Agentic Misalignment: How LLMs could be insider threats* (20 Jun 2025).
[71] J. Schlatter et al., *Shutdown Resistance in Large Language Models* (Palisade Research, Sep 2025).
[72] Y. Bengio et al, *International AI Safety Report 2026* (3 Feb 2026).
[73] United Nations, *Preliminary Report of the Independent International Scientific Panel on AI* (1 Jul 2026) p. 9.
[74] Centre for Long-Term Resilience, *AI loss of control incidents are worsening* (29 Aug 2026).
[75] W. Wang et al., *Let It Flow: Agentic Crafting on Rock and Roll, Building the ROME Model within an Open Agentic Learning Ecosystem* (31 Dec 2025).

Dates	Incident	How it was made public
11 May to 18 June 2026	**OpenAI:** during evaluations, OpenAI agents attacked a software service and uploaded hundreds of malicious packages.	Independent researchers identified the incident in September 2026.[76] OpenAI later confirmed its involvement.[77]
24 May to 22 June 2026	**OpenAI:** during evaluations, OpenAI agents hijacked a dormant German-language website to share over 18,000 messages with each other to find answers to test questions.	Independent researchers identified the incident in September 2026.[78] OpenAI later confirmed its involvement.[79]
9 to 13 July 2026	**OpenAI:** during evaluations, OpenAI agents breached testing environment restrictions and accessed Hugging Face's systems, contrary to their instructions.	Hugging Face disclosed that it was subject to a security incident on 16 July 2026.[80] OpenAI then confirmed its involvement and published its own disclosure five days later.[81]
April to July 2026	**Anthropic:** during evaluations, Claude AI agents breached third-party testing environments, accessed the internet and third-party systems. One agent published a malicious software package that was downloaded to 15 real systems.	Anthropic identified the incidents during a review of its evaluation transcripts, and disclosed it on 30 July 2026.[82]
July 2026	**Meta:** a misconfigured testing environment allowed a Meta model to access the internet. It exploited a vulnerability in a third-party company's service.	Journalists first reported on the breach on 5 August 2026.[83] Meta published its account on 14 August 2026.[84]
25 to 28 July 2026	**UK AISI:** during testing of Anthropic and OpenAI models, AI agents took a range of unapproved actions including trying to insert malicious code into open-source projects, social engineer people maintaining those projects, and contacting real people.	The UK AISI published an incident report on 4 August 2026.[85]

Australia has its own example of this risk materialising. When an Australian man asked his AI agent to book him a class at a gym, it did so by exploiting a vulnerability within the booking

[76] S. Kitts et al., *OpenAI agents carried out an undisclosed cyber-attack on RubyGems* (11 Sep 2026).
[77] *OpenAI agents attacked software service RubyGems before Hugging Face hack* (ABC News, 12 Sep 2026).
[78] S. Von Arx et al., *Discovery of a new OpenAI agent message board* (4 Sep 2026).
[79] A. Ha, *OpenAI confirms 'wiki incident', says it's working on a 'framework for more disclosure'* (TechCrunch, 5 Sep 2026).
[80] Hugging Face, *Security incident disclosure — July 2026* (16 Jul 2026).
[81] OpenAI, *OpenAI and Hugging Face partner to address security incident during model evaluation* (21 Jul 2026); METR, *Brief independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / Hugging Face hacking incident* (26 Aug 2026).
[82] Anthropic, *Investigating three real-world incidents in our cybersecurity evaluations* (30 Jul 2026).
[83] The Information, *Meta AI model hacked another company during cybersecurity testing* (5 Aug 2026).
[84] Meta, *Addressing an issue involving a third-party cyber evaluation of Muse Spark 1.1* (14 Aug 2026).
[85] UK AI Security Institute, *Incident Report: unsanctioned agent behaviour during cyber testing* (4 Aug 2026).

system to cancel another gym member's booking.[86] This example is relatively trivial, but shows how even when acting to achieve a legitimate goal instructed by a user, an AI agent can engage in inappropriate or unintended conduct. The same behaviour in a more sensitive setting, such as a hospital or financial institution, could cause significant levels of harm.[87]

This is not a solved problem

The frontier AI companies have no solution to the risk of loss of control or misalignment.[88] Independent assessments continually identify serious gaps in their current safety and control practices.[89] Commercial incentives are also making the problem harder. For example, OpenAI reported that its most recent release, GPT-6 Astra, "*shows a substantial decrease in chain-of-thought monitorability compared to previous models*".[90] This means its internal reasoning is less visible to developers and researchers, making it easier for the model to hide misalignment and misbehaviour.

Governments also lack visibility over this problem. Public knowledge of loss of control and misalignment incidents has depended on the goodwill of the frontier AI companies, the willingness of those impacted to disclose an incident, and work by independent researchers and journalists. This is not a credible mechanism for ensuring proper oversight of this high-risk technology. What is needed is a mandatory monitoring and reporting mechanism to ensure the government and regulators have proper visibility of the threat landscape.

Loss of control and misalignment are some of the most serious catastrophic risks that humanity faces from AI. The capability of frontier AI systems is advancing much faster than humanity's ability to understand and control them. Mitigating this risk requires significantly more investment in technical and governance measures that ensure we understand what these models are doing, can align them to human values, and maintain the ability to shut them down.

A.3 AI's potential to concentrate power

Finally, even in a world where the above AI harms above can be effectively managed, we will still need to contend with the potential for advanced AI models to concentrate political and economic power in a small number of countries and companies. The capabilities of these tools can also concentrate power within individual states, supporting authoritarian regimes that would otherwise lack the capabilities to exercise coercive control over their populations.[91]

Significant concentration is already seen across the AI supply chain. The Taiwan Semiconductor Manufacturing Company (**TSMC**) produces around 90% of advanced semiconductors.[92] Five

[86] C. Wilson, *AI assistant hacks gym website in first known Australian autonomous cyber attack* (ABC News, 9 Aug 2026).
[87] See, generally, Australian Signals Directorate, *Careful adoption of agentic AI services* (1 May 2026).
[88] See Z. Mowshowitz, *The Extinction Risk Preference Cascade: Quotes* (11 Sep 2026).
[89] Future of Life Institute, *AI Safety Index: Summer 2026* (7 Jul 2026); Guidelight AI Standards, *Guidelight's Control Assessment of Frontier AI Companies* (18 Aug 2026).
[90] OpenAI, *GPT-6 Astra System Card* (3 Sep 2026).
[91] See, e.g., E. Baptista and A.J. Vicens, *How Anthropic says Claude was used for weapons, spying, cyber operations* (Reuters, 11 Sep 2026), which reports on Mali using AI to build surveillance systems.
[92] R. Klingler-Vidra, *How Taiwan came to dominate the global chip industry* (The Conversation, 1 Apr 2026).

companies, Amazon, Microsoft, Google, Meta, and Oracle, control more than 70% of global AI compute.[93] Frontier and close-to-frontier models are all controlled by a few companies in the US and China.

The ability to determine access to frontier AI will provide significant leverage to specific governments and companies. Advanced AI will provide some governments with "*a country of geniuses in a datacenter*". [94] This new source of skills, expertise, and intelligence could result in significant strategic, economic, and military advantages. A Chatham House report warns that states that fail to secure influence over AI "*will likely forfeit control over their economies, societies, political systems and positions in the global economy*".[95]

This presents a real risk to Australia of losing key parts of its sovereignty and agency, and becoming economically disempowered. Australia's companies, government services, and critical infrastructure will become increasingly dependent on models not built, run in, or controlled by Australia. They will be governed by regulators overseas and terms set by foreign companies. As Prime Minister Albanese said, "*If we are always dependent on someone else, somewhere else, we will always be vulnerable.*"[96] Shadow Minister for Industry and Sovereign Capability, Andrew Hastie, has similarly warned of Australia becoming "*entirely dependent on other nations for our security*".[97]

We have already seen the warnings of what this potential dependent future could look like. In June 2026, the US government used export control powers to temporarily prevent non-US nationals from accessing Anthropic's most capable models.[98] We have also seen the US previously tightly regulate global access to frontier AI computing chips through so-called 'diffusion' rules.[99] AI companies are also voluntarily deciding to withhold model access, including from governments.[100]

Australia will have to accept a degree of dependence on AI systems built elsewhere. It is therefore critical that the government ensure that Australia has the assets to incentivise it being given continued access to those systems and that provide influence over the rules governing them. Those same assets will also provide resilience against disruptions, and enable stronger cooperation with other partners facing similar challenges.

---

[93] Epoch AI, *Hyperscalers control most compute* (14 Apr 2026).
[94] D. Amodei, *Statement from Dario Amodei on the Paris AI Action Summit* (Anthropic, 11 Feb 2025).
[95] F. Sandoval et al., *How Middle Powers Can Weather US and Chinese AI Dominance* (Chatham House, 16 Feb 2026).
[96] Prime Minister Anthony Albanese, *AI in Australia's interests* (15 Jul 2026).
[97] The Hon Mr Andrew Hastie MP, *Artificial Intelligence Joint Select Committee* (Hansard, 20 Aug 2026).
[98] J. Marinotti, *Legally or not, the US government is controlling global access to the world's most powerful AI* (The Conversation, 21 Jul 2026).
[99] T. Barrett & J. Lim, *The US AI Diffusion Rule: What is it, why did the United States rescind it, and implications for Australia* (United States Studies Centre, 21 Jul 2025).
[100] See, e.g., *The EU got access to Anthropic's most powerful model, three months later* (Euronews, 10 Sep 2026); *Anthropic Gives EU Access to Mythos Months After Model's Release* (Bloomberg, 10 Sep 2026).

B. Australia's agency in the AI era

While the challenges are significant, Australia is well placed to meet them. Doing so requires action on multiple fronts:

1. **First, building capabilities in key parts of the AI supply chain and demonstrating a value-add to allies and partners that cannot be easily replaced**.
2. **Second, setting precedents and norms domestically that others can follow, encouraging a 'race to the top' on AI security and governance.**
3. **Third, spearheading and supporting regional and international efforts to set the norms and write the rules on AI governance.**

B.1 Building domestic capabilities

It is not feasible for Australia, or indeed most countries, to be self-sufficient across the AI stack.[101] However, Australia does have the potential to develop capabilities in significant nodes of the AI supply chain. Building domestic capabilities that are valuable to partners would enable Australia to create interdependence with others that have capabilities in the rest of the AI stack, particularly the United States.[102] It could also provide us with greater optionality in relation to engaging with different AI models and suppliers.

Other middle powers have secured durable places in the AI supply chain. The Netherlands hosts a company that is critical to extreme ultraviolet lithography technology.[103] Japan is a major supplier of critical semiconductor inputs.[104] South Korean firms own a substantial portion of the computer memory market.[105] And Taiwan is well known for its chip industry.[106]

Australia does not yet have a comparable point of leverage. However, we have several potential opportunities to build these. This includes by:

1. **Building domestic AI compute capacity,** both for AI inference (i.e. serving AI models to customers) and AI training (i.e. the process that creates the AI model itself). Australia is highly competitive for data centre investment across energy, geopolitical risk, and geographic factors. However, reforms are needed to ensure safety and security are maintained, the sector's social license is real, and that other barriers such as copyright concerns are resolved fairly.

---

[101] H. Barasa et al., *Sovereignty in the Age of AI: Strategic Choices, Structural Dependencies and the Long Game Ahead* (Tony Blair Institute, 19 Jan 2026); Center for a New American Security, *Sovereign AI Index* (Aug 2026).
[102] B. Tanner et al., *Is AI Sovereignty Possible? Balancing Autonomy and Interdependence* (Brookings, Feb 2026).
[103] S. Van Der Lugt, *Why ASML is the EU's most Important Bargaining Chip* (Institute for European Policymaking, 22 May 2024).
[104] *Japan Country Commercial Guide: Semiconductors* (US ITC, 20 Nov 2025).
[105] Counterpoint Research, *Global DRAM and HBM Market Share: Quarterly* (Q2 2026).
[106] R. Klingler-Vidra, *How Taiwan came to dominate the global chip industry* (The Conversation, 1 Apr 2026).

2. **Supporting Australia's critical minerals industry**, in particular so that the refinement and processing of minerals important to AI hardware can occur onshore and in trusted partners. Australia has significant deposits of key minerals needed for the physical AI supply chain.[107] Realising the value of these endowments requires developing new markets for critical minerals exporters, expanding and deepening cooperation with partners to develop trusted supply chains, and reforms to support the industry's commercial sustainability.[108]

3. **Accelerating Australia's renewable energy build-out.** Global energy demand from data centres is projected to almost double to around 945 TWh over the next four years, and rise to 1200 TWh by 2035.[109] Time to power is also a key determinant in the attractiveness of sites for AI data centres.[110] Australia can help ensure that demand is filled with reliable, secure and clean energy options, while using AI investment to reach its clean energy goals.

4. **Investing in AI assurance.** As AI models become more capable and integrated into more sensitive systems, there will be increasing demand for services that evaluate, audit, and certify model safety, security, and reliability.[111] This includes software that tests AI systems against potential failure modes and hardware to verify compute is being used in accordance with contracted or regulatory requirements. Australia has an opportunity to build upon its expertise in other high-risk industries to develop businesses in this sector. This is also an area where focused attention by Australia's AISI could provide significant benefits, particularly in relation to hardware verification mechanisms.

Developing these assets would make Australia more useful to its partners, particularly the United States, helping to ensure it has access to frontier AI models and a voice in governance discussions.

This work should also be supported and lead to government-to-government agreements that provide certainty on the terms of Australia's access to frontier models and technologies. Options for this work include the United States—Australia Technology Prosperity Deal (currently under negotiation), the now more than 30 critical minerals arrangements Australia has with overseas partners, and plurilateral arrangements such as the Pax Silica initiative. Australia's trade agreements should also be reviewed and updated to support efforts to develop Australia's key AI assets and establish trusted AI supply chains.[112]

---

[107] See I. Satchwell, *Disruption and opportunity* (Australian Strategic Policy Institute, Feb 2026).
[108] See H. Channer & R. Monterosso, *Operationalising the Quad Critical Minerals Initiative* (United States Studies Centre, 8 Jul 2026).
[109] International Energy Agency, *Energy and AI* (Apr 2025).
[110] A. Phillips-Robins et al., *The Compute Coalition: How to Build the Future of AI in the Free World* (Carnegie Endowment for International Peace, 8 Jun 2026).
[111] See, generally, *AI Assurance Technology Report* (2024).
[112] Global Shield Australia, *Submission to the Department of Foreign Affairs and Trade's Southeast Asia Free Trade Agreement Modernisation Review* (31 December 2025).

Arrangements with private companies will also be key, including through contracts, government procurement,[113] and investment conditions imposed through the Foreign Investment Review Board process.

B.2 Setting precedents domestically

Australia has the opportunity to demonstrate regulatory success stories and use this to influence AI rules in jurisdictions around the world. Governments are actively grappling with how best to regulate AI, mitigate risks, and enable the benefits. In this context, early and credible movers can obtain disproportionate influence over how these rules evolve. Australia has already inspired technology regulation in other contexts, including regarding online harms.[114] The coming Australian Standards for AI are a further opportunity to replicate this in the context of AI data centre regulation and more broadly.

To successfully influence norms abroad, Australia needs a clear, consistent and coordinated approach to AI rules domestically. Global Shield Australia's recent *Guardrails & Gap Analyses* report identified 50 different processes relating to AI occurring at the Commonwealth level.[115] It recommended the Government provide greater transparency, urgency, and coordination to this work.

Framework or coordination legislation to underpin these processes would be a useful first step. Australia should learn from the approaches being taken in other jurisdictions and build upon elements that work for the Australian context, including risk management obligations, transparency requirements, and mandatory incident monitoring and reporting mechanisms.

B.3 Leading and supporting international efforts

Finally, Australia has the opportunity to multiply its influence through work in coalitions and with like-minded partners.

Australia has a history of leading international efforts to address global threats. Australia's establishment of the Australia Group helped to align export controls to help address chemical and biological weapon proliferation.[116] The Canberra Commission served as a key reference point for nuclear disarmament and non-proliferation.[117] More generally, Australia has regularly played a role in creating and sustaining international institutions and norms critical to international trade, peace and security. We should continue that tradition in the context of AI.

---

[113] A. Shamir, *Public procurement and AI governance: Developing a strategic framework for middle powers* (RAND, 11 Sep 2026).
[114] Department for Science, Innovation and Technology (UK), *Fact sheet: new rules to protect children online* (17 Jul 2026); *Malaysia enforces ban on social media accounts for children younger than 16* (NBC News, 1 Jun 2026); *New Zealand to introduce bill for under-16s social media ban* (BBC News, 24 Aug 2026).
[115] Global Shield Australia, *Guardrails & Gap Analyses: An Assessment of Australia's Approach to AI Regulation* (Aug 2026).
[116] Department of Foreign Affairs and Trade, *The Australia Group*.
[117] G. Evans, *The Canberra Commission On The Elimination Of Nuclear Weapons And Subsequent International Developments* (6 Oct 2000).

The National AI Plan commits Australia to using "*its role as a responsible middle-power to embed our values of safety, transparency and inclusion in international AI norms and standards.*"[118] Australia is currently participating in a range of international discussions on AI governance, including through the AI Summit series and the Hiroshima AI Process.[119] Australia also has AI and technology arrangements with a range of countries including Canada, India, Singapore, and the United Kingdom.[120] Several of Australia's trade agreements contemplate cooperation on AI,[121] and Australia is part of the United States' Pax Silica initiative.[122]

Other middle power democracies are already recognising the importance of cooperation and collaboration. Canada's *AI for All* strategy identified the importance of working with countries such as Australia, India, the United Kingdom, Japan and Germany.[123] Groups such as ASEAN are considering the region's approach to AI governance.[124] China has also established its own World AI Cooperation Organization, with 37 member states.[125]

Australia's yet to be released AI international engagement and regional leadership strategy must concretely set out the steps the government is taking to advance Australia's AI interests globally. This should include working with Pacific Island, Southeast Asian, and other partners to create institutions and mechanisms to support responsible development and uptake (for example, a regional AI Safety Institute hub or information and incident sharing arrangements). This should all be guided by clear statements recognising the catastrophic potential of AI, and a commitment to help set substantive norms, expectations, and red lines.[126]

C. What Australia can do in response

In rising to meet the challenges posed by AI, Australia is not starting from nothing. Processes are underway across the Commonwealth and the country to ensure regulatory frameworks are updated in light of the impact of AI.[127] In relation to catastrophic risks, this includes important reviews of AI's impact on synthetic biology, crisis management arrangements, and Australia's cyber and critical infrastructure laws. However, more could be done, and existing work needs to be completed far more quickly.

---

[118] Department of Industry, Science and Resources, *National AI Plan* (Dec 2025) p. 34.
[119] See, e.g., Department of Industry, Science and Resources, *Australia endorses India AI Impact Summit declaration* (Feb 2026); *Australia joins Hiroshima AI Process Friends Group* (2 May 2024).
[120] See Department of Industry, Science and Resources, *Overview of the memorandum of understanding to support the Australia-Canada-India Technology and Innovation (ACITI) Partnership* (10 Jul 2026); *Overview of the memorandum of understanding on artificial intelligence safety and security collaboration between Australia and the UK* (25 May 2026); *Memorandum of understanding on artificial intelligence safety collaboration between Australia and Canada* (5 Mar 2026); *Memorandum of understanding on cooperation on artificial intelligence between Australia and Singapore* (16 Dec 2024).
[121] Global Shield Australia, *Submission to the Department of Foreign Affairs and Trade's Southeast Asia Free Trade Agreement Modernisation Review* (31 Dec 2025) p. 7.
[122] United States Department of State, *Pax Silica Initiative* (Dec 2025).
[123] Politico, *Canada bids to lead middle powers in AI sovereignty race* (4 Jun 2026).
[124] S. Pillai, *The quiet regionalisation of AI governance* (Lowy Institute, 3 Sep 2026).
[125] Southward Tech, *WAICO signatories: 37 countries* (5 Aug 2026).
[126] See, e.g., The Future Society, *Global Red Lines for AI: A Three-Part Series* (3 July 2025); S. Russell et al., *AI governance through global red lines can help prevent unacceptable risks* (OECD GPAI, 22 Sep 2025).
[127] See Global Shield Australia, *Guardrails & Gap Analyses: An Assessment of Australia's Approach to AI Regulation* (Aug 2026).

More concretely, to successfully navigate a future with advanced AI, Australia needs to:

1. **Recognise the catastrophic risk** posed by advanced AI and that addressing this with international partners is a core national interest. This could be done in the Government's response to the Committee's report and in the coming Australian Government Strategy for International Engagement and Regional Leadership on AI.

2. **Harden Australia's defences,** including by mandating screening of genetic synthesis imports[128] and urgently completing the review of Australia's crisis management framework to prepare for AI incidents.

3. **Build Australia's contribution** to the AI supply chain (such as compute, critical minerals, clean energy, or AI assurance technologies), and use these to secure access to frontier models and influence on AI governance.

4. **Legislate to protect Australians**, joining like-minded partners and setting precedents for others to follow, including by:

    a. Enacting the 2027 Australian Standards for AI through a framework Act on AI that sets economy-wide definitions, standards, and rules, and addresses risks raised by frontier AI model developers, not just AI data centres. This should draw from regimes already established in California, New York, the European Union, Korea, Vietnam and elsewhere.

    b. Establishing a mandatory AI incident monitoring and reporting mechanism, modelled on systems overseas, to provide transparency of when AI models cause serious harms and near misses.[129]

    c. Using the coming reforms to the *Security of Critical Infrastructure Act* to incorporate AI data centres within Australia's critical infrastructure security frameworks and establish a path to also covering frontier AI models.

5. **Engage with partners** in the region and internationally to lead and support international AI governance. Australia should support major power cooperation to reduce AI's catastrophic potential and agree rules and norms around frontier AI. It should also work with neighbours in the Pacific and Southeast Asia, and partners further abroad, to build governance capacity, shared norms, and a common voice on key AI issues.

Decisions made now will determine whether Australia is able to flourish in a world changed by advanced AI, or is relegated to dependency on systems being developed elsewhere and rules being written by other governments. Australia has many options to shape its own AI future. It just needs political leadership that is willing to use them.

Contact: australia@globalshieldpolicy.org

[128] Australians for AI Safety, *Activate existing biosecurity powers to address AI-enabled risks* (4 May 2026).
[129] Global Shield Australia, *AI Incident Monitoring & Reporting for Australia* (Feb 2026).