

Turning AI Infrastructure into Benefits for Australia

Submission to the Office of AI's AI Standards Consultation Paper

October 2026

Submitted

9 October 2026

About Good Ancestors

Good Ancestors is an Australian charity dedicated to improving the long-term future of humanity by providing rigorous, evidence-based, and practical policy recommendations for Australia's biggest challenges. We have been deeply engaged in the AI policy conversation since our creation, working with experts around the world and helping to organise Australians for AI Safety.

Contact

If you would like to discuss this submission, please let us know at contact@goodancestors.org.au.



**Good
Ancestors**

Table of Contents

Summary of key points.....	3
Access to Frontier Models.....	5
Ecosystem and Skills Development.....	6
Co-locating research and engineering talent.....	6
Skilling Australians.....	7
Skilling for construction and operation.....	7
Skilling for AI competency and adoption.....	8
Public interest compute.....	9
Frontier AI Reporting Obligations.....	10
Who must provide information, and about what systems?.....	10
Cover AI companies carrying on business in Australia.....	10
Defining “frontier systems” through multiple tests.....	11
“Frontier systems” should be defined through a compute threshold.....	11
“Frontier systems” should also be defined through capabilities.....	12
“Frontier systems” should be defined through an explicit designation power.....	12
What information should be reported?.....	12
Companies should be required to report AI incidents.....	12
Capture incidents that occur during training, evaluation, and internal use.....	13
Capture “near miss” incidents.....	14
Set the relevant thresholds lower than international counterparts.....	14
Capture model theft or modifications.....	15
Companies should be required to report information about AI capabilities.....	15
Cover internal models.....	15
Include detailed research and development metrics.....	16
Include detailed biosecurity capability evaluations for all models.....	16
Including information already disclosed to overseas regulators.....	17
Risk Management Obligations.....	18
Australia’s crisis management and critical infrastructure laws.....	18
AI in the Australian Government Crisis Management Framework.....	18
AI in the Security of Critical Infrastructure Act.....	18
Compute verification obligations.....	19
Why is compute verification important?.....	19
What might compute verification obligations look like?.....	19
Embedded safety evaluators.....	20
What might embedded safety evaluators look like?.....	20
Conclusion.....	21

AI Standards Consultation

Australia should host frontier AI training with strings attached. Few countries will host training. Australia has energy, space, and stability that few others can match. However, while Australia has a strong hand, we should not overplay it. Frontier AI companies have genuine red lines and will balance the overall Australian operating picture with their best available alternatives.

The consultation paper gets the foundations right. Resolving the questions in Part 2 of the *Consultation Paper* is essential to securing social licence and ensuring AI data centres are consistent with Australian wellbeing. Our submission to the *Senate Inquiry into AI and Data Centres* recommended mandatory standards tackling genuine issues around power, water and community impacts.¹ In February 2026, Anthropic committed to cover 100% of the grid upgrade costs of its data centres and to offset consumer price increases. Other companies have made similar pledges.²

Enforcing these commitments will benefit Australians, but they are not the only form of conditions Australia should consider. This submission responds primarily to Part 3 of the *Consultation Paper*.³ We discuss four areas:

1. Access to frontier models for Australian businesses, critical infrastructure operators, national security agencies, and the AI Safety Institute
2. Obligations to build Australian skills and public-interest compute
3. Mandatory incident and transparency reporting, and
4. Participation in compute verification and emergency management

Government should aim to gain the most value it can from frontier AI companies without those companies moving to a different jurisdiction to train frontier models.

Any overall deal could include many elements. Some are red lines for AI companies. If we cannot solve the copyright issue, frontier AI training cannot happen in Australia. Other issues, like securing access to frontier models, should be top priorities for Australia. This submission details a range of elements, but we recommend against Government including every conceivable element because it would create unnecessary complexity. We recommend Government prioritise commitments that cannot be purchased on the free market and pursue fees and taxes to fund other interventions.

Frontier model access falls into the first category. By definition, pre-release model access is not available on the market. Public-interest compute and skills lean towards the second category.

Summary of key points

- Australia should host frontier AI training with strings attached. We should seek the most value we can without pushing frontier AI companies to train elsewhere.
- Frontier AI companies training here should guarantee Australia the same model access as the US for the public, critical infrastructure operators, and national security agencies.

¹ Good Ancestors. (2026). [Hosting AI infrastructure offers Australia more than economic benefits.](#)

² Anthropic. (2026). [Covering electricity price increases from our data centers.](#)

³ Department of the Prime Minister and Cabinet. (2026). [Getting it right: Building AI infrastructure that works for Australia](#); White House Office. (2026). [President Trump secures historic commitment to keep electricity costs down amid data center boom.](#)

- Model access needs to be secured *both* through frontier AI companies and the US Government. Linking Australian model access to permission to train in Australia incentivises US companies to encourage the US Government to preserve Australian access.
- Hosting frontier AI should build Australian skills and compute capacity, with obligations that scale with the amount of compute a company hosts here. This could include:
 - Requiring frontier AI companies training here to keep a research and engineering team in Australia whose principal work is developing, evaluating, or securing frontier models.
 - Including proportionate skills commitments in mandatory standards for data centres during construction and operation, such as apprenticeship targets and TAFE partnerships.
 - Requiring frontier AI companies to fund 'AI apprenticeships', modelled on Anthropic's Claude Corps, that place early-career Australians in not-for-profits, public research organisations, and government agencies.
 - Requiring large compute operators to 'reserve' a generous share of their Australian-based capacity and offer it to Australian buyers first, at market rates.
 - Requiring companies training AI in Australia to gift a portion of their Australian training capacity to not-for-profits, researchers, start-ups, and AI Safety Institute evaluations.
- Australia should impose robust transparency and incident-reporting obligations on AI companies that carry on business in Australia.
 - Incident reporting should set lower thresholds than equivalent laws in other jurisdictions, cover internal use, near misses, misaligned behaviours, and theft of model weights.
 - Transparency reporting should cover internal models, detailed AI R&D automation evaluations, and full biosecurity evaluations. Australia will have a stronger regulatory basis if the model is trained here.
 - Frontier AI companies should be required to share reports given to overseas regulators.
- Hosting training brings frontier AI companies within reach of Australia's crisis management and critical infrastructure laws. Those laws need updating to ensure Australia benefits.
 - Update the Australian Government Crisis Management Framework and related mechanisms like the National Coordination Mechanism and the Trusted Information Sharing Network to acknowledge and prepare for AI-created and AI-mediated risks.
 - The *Security of Critical Infrastructure Act* should capture entities that train AI (not only data centre operators), and reach AI providers that other critical infrastructure depends on.
- Frontier AI companies training here should work with the AI Safety Institute to build compute verification tools, and in the long term, be required to use them in their Australian facilities.
- Australia should require embedded third-party safety evaluators in frontier AI companies—a measure which Anthropic and OpenAI have already called for.

Access and Capability Conditions

Access to Frontier Models

Australia must secure ongoing access to frontier models. This may be the most important term we negotiate from hosting training. On September 18, the Director-General of the Australian Signals Directorate said:⁴

Assurance of ongoing access will become increasingly important as critical functions, such as defence, intelligence and financial systems, are increasingly dependent on AI ... Access to frontier models is a matter of urgency.

Losing access to frontier AI has implications for national security, the economy, and research. National security can be asymmetric. In the cyber domain, attacks need to find a single vulnerability, while defenders have to secure everything. Policy should help defenders gain earlier access to capable AI while delaying and suppressing attackers' access and capability. Policy should give defenders as much time and capability advantage as possible to secure critical infrastructure and systems of national significance.

Businesses with more capable AI models may outcompete those only using less capable models. Particularly in globally competitive fields, an economic gap could grow between AI 'haves' and 'have-less' or 'have-nots'.

The Australian AI Safety Institute (AISI) needs frontier-model access to evaluate capabilities and risks, and Universities and research institutions need access to do high-quality research and build local expertise.

Ongoing access to frontier AI is not assured. Australia has already struggled to gain access to advanced models.⁵ Even the UK's AI Security Institute did not secure pre-release access to Mythos 5.1.⁶ Frontier access may become harder as security concerns rise and early access becomes more valuable. Currently, Australian early access depends on goodwill, not the real value our access provides to our partners.

Australia should seek the same access to frontier models that the US receives at three levels:

1. Market access for the Australian public, equivalent to the US public,
2. Access to restricted models for critical infrastructure operators, equivalent to their US counterparts
3. Government access for national security purposes, equivalent to the US Government.

Government should negotiate with both AI companies and the US Government.

- **Negotiations with AI companies.** Companies training in Australia could be obligated to provide model access on par with the United States.⁷ One way to enforce this is by linking Australian model access to permission to train in Australia. This directly obliges companies and incentivises them to encourage the US Government to preserve Australian access.
- **Negotiations with the US Government.** National governments have various powers to override commercial decisions by companies. This means developer obligations need to be paired with agreements with relevant national governments.

⁴ Australian Signals Directorate. (2026). *Evidence to the Joint Select Committee on Artificial Intelligence*, Proof Committee Hansard.

⁵ Anthropic's Mythos cybersecurity model was initially only released to a restricted group of mostly US-based organisations. This did not include the Australian Government. When a public version was released (Fable 5), it was available for only three days before a US Government export directive required Anthropic to suspend access for foreign nationals.

⁶ Kelly, R. (2026). [Anthropic reportedly withholds access to Mythos 5.1 from UK safety testing body](#). IT Pro.

⁷ Leicht, A. (2026). [Compute for Access](#). Threading the Needle.

Securing frontier access is easier if our access creates value for the countries and companies we partner with. If our AISI conducts valuable research that AI companies want, they have more reason to give us access. If our national security agencies use access to secure the critical infrastructure that powers our partners' compute, they have more reason to give us access. If our researchers and businesses are part of intertwined supply chains with their researchers and businesses, they have more reason to give us access.

Ecosystem and Skills Development

The paper asks how frontier AI companies should contribute to Australian capability. We propose three conditions:

1. AI companies locate research and engineering staff in Australia
2. "AI apprenticeships" that train early-career Australians and deploy them into Australian workplaces, and
3. A national compute reserve and donated public-interest use.

Co-locating research and engineering talent

An Australia-based training cluster could be operated from overseas. However, co-location is important for ecosystem spillover benefits for Australian universities, start-ups, the AI Safety Institute, and other elements of the Australian AI supply chain. To address this, AI companies training AI in Australia should be required to locate research and engineering staff in Australia. Australia should focus on attracting the highly skilled researchers and engineers who build, evaluate, and secure AI models.

A headquarters that functions as a marketing and engagement hub also doesn't achieve spillover benefits, and data centres themselves may employ few people in ongoing positions. Virginia's Joint Legislative Audit and Review Commission found in 2024 that a typical 250,000 square foot data centre employs about 50 people, half of them contractors, and about 1,500 construction workers at the peak of the build.⁸

Other countries have used similar conditions, for example:

- Nigeria's 2010 content law caps expatriates at 5% of an operator's management positions, requires positions to pass to Nigerians after at most four years of understudy, and sets minimum Nigerian content by activity.⁹
- Ireland refuses employment permits to a firm unless at least 50% of its staff are EEA nationals. The rule is waived for two years for start-ups backed by IDA Ireland or Enterprise Ireland.¹⁰
- Saudi Arabia's regional headquarters licence requires 15 full-time staff, three of them senior executives, within a year. From 1 January 2024 the Saudi Arabian Government ceased contracting with multinationals without regional headquarters.¹¹
- From 1979 to 1994, Norway required oil field operators, Norwegian and foreign, to do at least 50% of the research needed to develop a field in Norway. Foreign companies also signed goodwill agreements to do research in Norway, and the Government weighed these in licence awards. Foreign

⁸ Joint Legislative Audit and Review Commission. (2024). [Data Centers in Virginia](#). Commonwealth of Virginia.

⁹ [Nigerian Oil and Gas Industry Content Development Act 2010](#) (Nigeria), s 32 and Schedule; KPMG Nigeria. (2010). [Commentary on the Nigerian Oil and Gas Industry Content Development Act, 2010](#)

¹⁰ Department of Enterprise, Tourism and Employment. [General Employment Permit](#). Government of Ireland.

¹¹ Vistra. (2024). [What you need to know about Saudi Arabia's local headquarters rule](#); Sovereign Group. (2024). [Saudi issues new guidance for Regional Headquarters Programme](#).

companies devoted more than NOK 10 billion to research, product development, and skills in Norway over that period.¹²

- Brazil requires operators of high-productivity oil fields to spend 1% of gross revenue on research and innovation. The original clause sent at least 50% of that to Brazilian universities and research institutes accredited by the regulator. Contracts awarded since 2017 require 30–40% for those institutions and a further 30–40% for innovation projects with Brazilian companies. An evaluation based on 73 interviews and a survey of 156 project coordinators found the clause strengthened Brazilian research and technological capability.¹³

We propose each AI developer maintain a research and engineering team in Australia whose principal work is developing, evaluating, or securing frontier models. As a benchmark, a developer's share of technical staff in Australia should match its share of training compute in Australia.

For example, a developer with 20% of its training compute in Australia should base 20% of its research and engineering staff in Australia. The developer would agree a binding workforce plan with the Government, setting out the full-time-equivalent positions and capabilities of its Australian team. This lets frontier AI companies scale their Australian compute while keeping a proportionate number of research and engineering positions in Australia.

Skilling Australians

Government can deliver positive workforce development outcomes via standards on frontier AI companies, compute providers, and data centre operators. The *consultation paper* separates requirements for large data centres and conditions for AI training into separate parts.

Skilling for construction and operation

Standards that uplift Australia's digital infrastructure workforces benefit frontier AI companies and Australians. A workforce capable of building and operating digital infrastructure at greater scale increases Australia's attractiveness as a destination for AI investment. That investment can, in turn, support further workforce development, creating a cycle that helps Australia capture more benefits from AI.

A compounding cycle of investment and workforce capability would help Australia build a globally significant amount of compute. Compute capacity that meets domestic demand, supports our region, and contributes meaningfully to frontier AI training is strategically significant and advances the national security interests of Australia and our partners.

Skills development opportunities from AI infrastructure extend across the technology stack, from civil construction, to operating computer networks and systems, developing, deploying and evaluating AI systems. Standards could support apprenticeships and placements in electrical trades, civil engineering, or cooling systems management. Post-construction data centre operation can support ongoing capability uplift in networking and cybersecurity. Frontier AI companies undertaking training here could contribute to Australian expertise in AI model development, research, and safety evaluation.

Globally, countries are navigating the question of how best to capture these benefits in different ways:

¹² Ryggvik, H. (2015). [A Short History of the Norwegian Oil Industry: From Protected National Champions to Internationally Competitive Multinationals](#). Business History Review; Norwegian Petroleum Museum. [Draugen: Goodwill agreements](#); Columbia Center on Sustainable Investment. (2016). [Local Content: Norway Petroleum](#).

¹³ Mayer Brown. (2022). [Research, Development and Innovation in the O&G Industry](#). Brazil Energy Journal; Mancini, L. & Paz, M. J. (2018). [Oil sector and technological development: Effects of the mandatory research and development \(R&D\) investment clause on oil companies in Brazil](#). Resources Policy.

- Thailand, an emerging market for data centres in the Asia-Pacific, requires eligible data centre and cloud service developers to submit and implement a plan delivering local benefits to Thailand.¹⁴ Activities can include professional training and skills development, partnerships with educational institutions, and local procurement.
- Canada introduced its Responsible Data Centre Development Principles in September 2026, which offers general expectations that data centres provide lasting local benefits to communities, including skills training, apprenticeships, and research partnerships.¹⁵ Canada's approach is government-led, non-binding and industry-endorsed.¹⁶
- A June 2025 MoU between the UK and NVIDIA outlines points of collaboration for AI skills in AI model development and deployment.¹⁷ It proposes hands-on training for students, researchers and technical professionals through labs and sandbox environments, and access to NVIDIA's Deep Learning Institute for accredited educational institutions.

Workforce obligations should reflect the different skills needed for construction and operation.

Data centre builders could meet skills and training obligations drawing on established requirements for government projects, like the Australian Skills Guarantee targets for apprentices and trainees. New requirements could also be developed requiring engineering placements or partnerships with registered training organisations. Obligations during this phase should nurture skills relevant to data centre construction and fit-out, expanding the skilled workforce available to deliver future projects at greater scale.

Ongoing data centre operation could support technical placements and training by compute providers in systems administration, networking and cybersecurity. Data centre operators could support supervised placements and ongoing training in power management, equipment fault diagnosis, or facility management. This would grow the workforce needed to operate data centres at scale.

Australia's data centre development pipeline is substantial, valued at \$29.4 billion in New South Wales alone.¹⁸ Investment in digital infrastructure is expected to keep growing. Meeting demand for construction and ongoing operations will require our workforce to operate at a greater scale. Government may consider further initiatives to support workforce scaling, such as common competency and assessment standards for data centre operations. These could be incorporated into a nationally recognised training framework, enabling credential portability across operators and jurisdictions.¹⁹

Skilling for AI competency and adoption

Beyond the skills needed to build and operate AI infrastructure, AI companies could fund 'AI apprenticeships' that develop practical AI skills while strengthening public-interest organisations.

Claude Corps, launched in June 2026, will pay 1,000 early-career Americans a US\$85,000 salary to spend a year embedded in more than 400 not-for-profits, backed by an initial US\$150 million. Anthropic says it wants a model it can replicate in other countries.²⁰ Scaled by population, the US program equals about 80

¹⁴ Office of the Board of Investment (Thailand). (2026). [Investment Promotion Guide 2026](#).

¹⁵ Government of Canada. (2026). [Canada's Responsible Data Centre Development Principles](#). Innovation, Science and Economic Development Canada.

¹⁶ Government of Canada. (2026). [Canada's Responsible Data Centre Development Principles](#). Innovation, Science and Economic Development Canada.

¹⁷ UK Department for Science, Innovation and Technology & NVIDIA. (2025). [Memorandum of Understanding between UK and NVIDIA on AI skills](#).

¹⁸ Future Skills Organisation and Powering Skills Organisation. (2026). [Submission to the NSW Legislative Council Public Accountability and Works Committee Inquiry into Data Centres](#). Submission No. 118, NSW Parliament.

¹⁹ Future Skills Organisation and Powering Skills Organisation. (2026). [Submission to the NSW Legislative Council Public Accountability and Works Committee Inquiry into Data Centres](#). Submission No. 118, NSW Parliament.

²⁰ Anthropic. (2026). [Introducing Claude Corps](#).

Australian fellows a year.²¹ These 'AI apprentices' could be placed in Australian not-for-profits, public research organisations, and government agencies, with the AI developer paying salaries and providing model access. Further cooperative conditions could be applied which build the capabilities of Australian personnel at the AISI in safety evaluation or verification. This would build a pipeline of individuals with AI skills and help recipient organisations upskill.

Public interest compute

Australia has a significant compute deficit, which is forecast to worsen over time. We estimate Australia hosts about 0.4% of global AI compute, enough to serve about 39% of our own inference demand, falling to under a quarter by the end of 2027 on announced builds.²²

Globally, inference supply is growing about 3 times a year, while demand is growing about 10 times a year. These trends lead experts to forecast a "compute crunch".²³ When compute is short, operators may serve their home market and their largest customers first.

Australian policy should avoid a situation where Australia hosts significant AI compute, but it's purchased by overseas customers and not accessible by Australians. Anticipating this, Australian data centres hosting compute could be required to reserve capacity for Australia under two mechanisms, similar to the consultation paper's "compute reservation scheme" and "public-interest research payment".²⁴ There are two separate but related challenges:

- **Availability:** whether Australian buyers can access scarce compute. This could be addressed by offering compute to the Australian market first. This ensures access for Australians while imposing little regulatory burden.
- **Affordability:** whether Australian not-for-profits, research institutes, the AI Safety Institute or start-ups can afford compute. This could be addressed by donating compute. This ensures some access for relevant Australian institutions, albeit with more regulatory burden.

This challenge has been navigated by other countries and industries.

- Western Australia has required LNG projects since 2006 to reserve gas equivalent to 15% of production for the domestic market, as a condition of access to State land for processing facilities.²⁵ Domestic buyers pay for the gas. Commitments are written into agreements for each project. A 2024 WA parliamentary inquiry found the policy no longer fit for purpose, citing a lack of consistency, transparency, and enforceability.²⁶ The policy was updated in September 2024 to focus on timely delivery.²⁷
- The Commonwealth's *Australian Domestic Gas Security Mechanism*, in place since 2017, lets the Minister restrict LNG exports when a domestic shortfall is forecast.²⁸ It has never been activated. The prospect of it produced supply agreements with exporters instead.²⁹

²¹ 1,000 fellows × Australia's population (27.6 million) ÷ the US population (341.8 million) ≈ 80.

²² Country shares are from Epoch AI's GPU cluster database (epoch.ai/data/gpu-clusters), which covers an estimated 10–20% of global compute. Epoch considers the distribution broadly representative. We apply those shares to Epoch's estimate of approximately 30 GW of total global AI data centre power capacity in late 2025.

²³ Emberson, L. & Sevilla, J. (2026). [Is a compute crunch coming?](#) Epoch AI.

²⁴ Department of the Prime Minister and Cabinet. (2026). [Getting it right: Building AI infrastructure that works for Australia](#);

²⁵ Government of Western Australia. (2026). [WA Domestic Gas Policy](#).

²⁶ Economics and Industry Standing Committee. (2024). [Domestic Gas Security in a Changing World, Inquiry into the WA Domestic Gas Policy: Final Report](#). Legislative Assembly of Western Australia.

²⁷ Government of Western Australia. (2024). [Updated Domestic Gas Policy to Secure WA's Energy Future](#).

²⁸ Department of Climate Change, Energy, the Environment and Water. (2025). [Gas Markets](#). Australian Government.

²⁹ Department of Industry, Innovation and Science. (2020). [Review of the Australian Domestic Gas Security Mechanism](#). Australian Government.

- The United States' National AI Research Resource pilot pools about US\$100 million of donated compute and model access from 28 non-government companies, including US\$20 million of Azure credits from Microsoft and about US\$30 million from Nvidia. It supported more than 600 projects and 6,000 students in its first two years.³⁰

Australia could adopt two policies:

- **National priority access.** Operators above a certain threshold (perhaps 100 MW of aggregate AI compute in Australia), whether frontier AI companies, hyperscalers, or neoclouds, should offer 15% of that capacity to Australian buyers first, at market rates. The price should be no higher than the operator's best rate to a comparable customer.
- **Public-interest compute.** Operators above a certain threshold (perhaps 100 MW of aggregate AI compute in Australia) should donate 1% of their aggregate Australian capacity to a pool for not-for-profits, research institutes, and early-stage start-ups, with a dedicated tranche for the AI Safety Institute.

Public-interest compute should strengthen Australia's domestic AI and compute capability without crowding out local providers. Given the value judgments involved in deciding which causes or sectors government should support, the best way to achieve this may be either to allocate public-interest compute resources directly to AISI, allocate via the National Computational Merit Allocation Scheme (NCMAS), or via a new mechanism modelled on the NCMAS.

The NCMAS sets compute priorities and selects recipients using a transparent and independent criteria-based process. Projects that do not receive allocations may not proceed, rather than defaulting to purchasing compute on the open market, meaning that compute distributed this way does not crowd out local providers.

Frontier AI Reporting Obligations

Regulators must know what frontier AI systems can do, the risks they pose, and how companies are reducing those risks. Currently, Australian regulators only know what is public. Publicly available information about publicly available models is not good enough. The Australian Signals Directorate has called for laws that require frontier AI companies to share testing results, information on unexpected model behaviour, and information on misuse.³¹

Who must provide information, and about what systems?

Cover AI companies carrying on business in Australia

A developer that sells frontier AI models to Australians should report to Australian regulators, whether or not it trains here, or is based here. Other goods supplied in Australia must meet Australia's mandatory safety standards, wherever they are made. AI should work the same way.

Whether an offshore company "carries on business in Australia" is an established test for who we can regulate. Using this test for AI would align with the *Privacy Act 1988* and the *Competition and Consumer Act 2010*. Such a test would cover companies offering AI services to Australian consumers and those training their AI products in Australia.³²

³⁰ U.S. National Science Foundation. (2024). [National Artificial Intelligence Research Resource](#).

³¹ Australian Signals Directorate. (2026). Evidence to the Joint Select Committee on Artificial Intelligence, Proof Committee Hansard.

³² See, eg, *Clearview AI Inc and Australian Information Commissioner* [2023] AATA 1069; *Valve Corporation v Australian Competition and Consumer Commission* [2017] FCAFC 224.

Defining “frontier systems” through multiple tests

Any incident and transparency reporting scheme should:

- Ensure the category of models the scheme applies to – what we call “frontier systems” – captures the forms of AI that raise the most serious risks, and to the extent possible, does not capture non-risky models.
- Ensures the boundaries of the relevant categories are flexible, to allow regulation to adapt to technological change.

In light of these goals, we suggest that any Australian regulatory scheme should define “frontier systems” through three different paths:

- A classification as “frontier systems” by reference to a compute-threshold – changeable through delegated legislation or rules.
- A classification as a “frontier systems” by reference to a selection of capability improvements.
- A separate pathway, set by ministerial designation, if the above pathways are insufficient.

“Frontier systems” should be defined through a compute threshold

Overseas legislation largely defines “frontier models” by reference to floating-point operations, or ‘FLOPs’.

Jurisdiction	Compute threshold	Other routes?	Alteration to threshold?
EU AI Act	10 ²⁵ FLOPs of training compute, including fine-tuning and later reinforcement learning. ³³	The Commission can designate a model, on its own initiative or after an alert from its scientific panel, if a model's capabilities or impact match the most advanced models. ³⁴	EU Commission. ³⁵
Californian Transparency in Frontier Artificial Intelligence Act (TFAIA)	10 ²⁶ FLOP of training compute, including fine-tuning and later reinforcement learning. ³⁶	No. Compute is the only test for being classified as a ‘frontier model’. ³⁷	Legislative alteration required. ³⁸

The differences are:

- The EU AI Act has a compute threshold that is 10x lower than the Californian TFAIA.
- The Californian TFAIA thresholds can only be changed via legislative amendment, but the EU AI Act thresholds can be changed by administrative action.

Australian regulation should have flexibility. Algorithmic improvements mean less compute is needed for the same capabilities each year,³⁹ so fixed thresholds quickly go out of date. Regulators should be able to update it without waiting for Parliament.

³³ [Regulation \(EU\) 2024/1689 \(Artificial Intelligence Act\)](#) (European Union), art 51(2), recital 111.

³⁴ [Regulation \(EU\) 2024/1689 \(Artificial Intelligence Act\)](#) (European Union), arts 51(1)(b), 52(4), Annex XIII.

³⁵ [Regulation \(EU\) 2024/1689 \(Artificial Intelligence Act\)](#) (European Union), arts 51(3), 97.

³⁶ [Transparency in Frontier Artificial Intelligence Act 2025](#) (California), s 22757.11(i).

³⁷ [Transparency in Frontier Artificial Intelligence Act 2025](#) (California), s 22757.11(h)–(i).

³⁸ [Transparency in Frontier Artificial Intelligence Act 2025](#) (California), s 22757.14.

³⁹ Epoch AI (2024), [Algorithmic Progress in Language Models](#).

Good Ancestors has no position on whether the initial threshold should be 10^{25} or 10^{26} . EpochAI suggests that many models have already been trained that exceed the EU AI Act's threshold.⁴⁰ So, the lower the threshold, the more resources regulators should have to process and investigate all reports.

"Frontier systems" should also be defined through capabilities

Compute thresholds alone are not enough. Algorithmic improvements mean models trained with less compute can match the capabilities of today's largest models.⁴¹ Tool access and multi-agent setups can also make a system more capable than its training compute suggests.⁴² A scheme that relies only on compute will miss some dangerous systems, and miss more each year.

Australia's scheme should therefore also capture systems by what they can do. A system below the compute threshold should still be a frontier system if, when used with tools, scaffolding, harnesses or other agents, it shows capabilities comparable to models above the threshold. The regulator, advised by the Australian AI Safety Institute, should set the relevant capabilities and benchmarks in rules it can update.

This means that frontier AI companies hold the information needed to tell whether a system meets the test. We do not resolve this here, but it is one reason independent oversight, such as embedded safety evaluators (discussed below) is an important part of AI governance.

"Frontier systems" should be defined through an explicit designation power

Given technology is changing rapidly, even compute and capability tests may miss risky AI. A designation power lets government bring a specific model or system into scope straight away if it receives information that it likely poses serious risks.

What information should be reported?

Regulators need two kinds of information. Transparency reporting requires AI companies to tell regulators what their frontier AI models can do. This helps regulators anticipate risks. Incident reporting tells them, within a fixed time of an incident, when something has gone wrong. This helps regulators respond to failures and learn from them.

Australian policymakers distinguish between two horizons of AI safety.⁴³ The first covers harms that are real and present, such as biased automated decisions and chatbots that harm children.⁴⁴ The second covers novel frontier risks that could be catastrophic and irreversible, such as AI helping someone build a bioweapon.

This section focuses on the second horizon. Regulators don't have enough information about these risks, and need to know about relevant frontier incidents and capabilities to measure them. First-horizon harms are serious and need strong regulation but are best addressed through existing mechanisms, even if those mechanisms need to be strengthened.

Companies should be required to report AI incidents

AI companies should tell government when a reportable incident occurs. They should share technical details about how the incident occurred, the model involved and the type of harm created, technical details about

⁴⁰ Epoch AI. (2025). [Over 30 AI models have been trained at the scale of GPT-4.](#)

⁴¹ Epoch AI (2024), [Algorithmic Progress in Language Models.](#)

⁴² Reid, A. et al. (2026). [Risks and controls for multi-agent systems: An analytical framework for deployment of AI agents across organisational boundaries.](#) Australian AI Safety Institute.

⁴³ Charlton, A. (2026). [AI Safety Forum.](#) Speech, Sydney, 7 July.

⁴⁴ Charlton, A. (2026). [AI Safety Forum.](#) Speech, Sydney, 7 July.

the incident, why safeguards failed, the people involved and affected, and the relevant response by the developer.⁴⁵

The *EU AI Act* and *California's Transparency in Frontier Artificial Intelligence Act* have the most developed incident reporting schemes. Australia should draw on their structure, deadlines and report content. But they both define a reportable incident too narrowly.

- The thresholds are too high. Both schemes generally require death, serious injury, or other substantial harms before an incident is reportable.⁴⁶
- Apart from some narrow exceptions, neither scheme requires a report when model safeguards fail, but no harm actually occurs.⁴⁷
- They do not adequately capture stealing or unauthorised modification of frontier models.
- They do not reliably and adequately cover training, evaluation and internal use.⁴⁸

Any Australian scheme should close these gaps. Each is discussed below.

Capture incidents that occur during training, evaluation, and internal use

Any incident reporting scheme must clearly and explicitly cover *internally-deployed* frontier AI—frontier AI models not available to the public—because:

- **Internal models pose significant risks.** The HuggingFace incident showed how an internal model can cause harm by pursuing a goal in ways its developers didn't intend. Internal models are often more capable than public models, meaning they may carry greater risks but less scrutiny.
- **Internally deployed systems operate with reduced safeguards.** Internal models often lack the safeguards applied to public models.⁴⁹ An employee could use internal models to cause serious harm, such as by developing a bioweapon.⁵⁰
- **Current internal models may eventually be made public.** Today's internal models are tomorrow's public models. Information about internal deployment gives governments time to prepare for future risks.

Covering internal models expressly would make Australia a world leader. Other countries' regulation does not adequately cover internal models.

- The EU AI Act's coverage of internal models remains legally uncertain, and its research and development exemptions likely leaves important internal deployments outside its scope.⁵¹
- *California's Transparency in Frontier Artificial Intelligence Act* does include some reporting requirements on "internal frontier models" but these lack the comprehensiveness of the *EU AI Act's* market-model reporting requirements.⁵²

⁴⁵ See, eg, Dixon, R. B. L. & Frase, H. (2025). [AI incidents: Key components for a mandatory reporting regime](#). Center for Security and Emerging Technology.; European Commission. (2025). [Code of Practice for General-Purpose AI Models: Safety and Security Chapter](#).

⁴⁶ [Regulation \(EU\) 2024/1689 \(Artificial Intelligence Act\)](#) (European Union), art 3(49); [Transparency in Frontier Artificial Intelligence Act 2025](#) (California), s 22757.11(c)–(d).

⁴⁷ Kara, Z. (2026). [Improving Frontier AI Incident Reporting Regimes](#). Centre for the Governance of AI.

⁴⁸ [Regulation \(EU\) 2024/1689 \(Artificial Intelligence Act\)](#) (European Union), art 2(6), (8). The Californian Act does cover internal models, but has thresholds that make the coverage ineffective. This is discussed further below.

⁴⁹ Kwon, J. & Casper, S. (2026). [Internal deployment gaps in AI regulation](#). arXiv.

⁵⁰ Delaney, O. et al. (2026). [Risk reporting for developers' internal AI model use](#). Institute for AI Policy and Strategy.

⁵¹ The precise legal position is complex. For discussion, see Pistillo, M. (2026). [Internal deployment in the AI Act](#). Cambridge Commentary on EU General-Purpose AI Law.

⁵² See further, Charnock, J. et al. (2026). [What should frontier AI developers disclose about internal deployments?](#) arXiv.

Capture “near miss” incidents

Australia's scheme should require frontier AI companies to report near misses. Near misses provide valuable information. They show where safeguards are failing while there is still time to fix them.

A near miss is an unplanned event or system behaviour that did not cause harm but could have. Classifying an event as a near miss “hinges on whether there was a credible path to serious harm before being interrupted”.⁵³ For example, companies should be required to report incidents where a model “attempts an action capable of causing serious harm but is stopped by a last-minute human intervention”.⁵⁴ This helps regulators understand which safeguards are working, where serious harm is being prevented by only a single safeguard, and how common future incidents might be.

Broader reporting should give the regulator useful warnings without burying them in low-value reports or reports that existing regulators already handle. Drawing this line will be hard, but policymakers can use techniques from other fields like aviation safety and therapeutic goods.⁵⁵ For example, whether an aviation occurrence must be reported immediately, reported in writing within 72 hours, or not reported at all depends on the type of occurrence and the type of operation.⁵⁶

Set the relevant thresholds lower than international counterparts

Australia should set lower harm-thresholds than the EU and California. It should also consider adding a second type of reportable incident based on what a system does, even if it causes no harm:

- **Lower outcome thresholds.** Any death or injury, cyber intrusion, material financial or property damage, or disruption to critical infrastructure should be reportable, whether or not it is irreversible, catastrophic, or very serious in scale.
- **A list of reportable behaviours.** Some behaviours should be reportable whatever their outcome. Examples include a model accessing systems it is not authorised to use, attempting to evade monitoring, copying its own weights, or expressly deceiving its developer.

Serious harm is rare. A scheme that waits for serious harm gives regulators little to learn from until it is too late. Californian officials said OpenAI agents' breach of Hugging Face did not meet their reporting thresholds.⁵⁷

The precise thresholds, and which behaviours go on the list, should be worked out with the Australian AI Safety Institute, frontier companies, and relevant regulators. They should be set in rules the regulator can update as risks change.

Reportable behaviours are specific actions by a model that indicate misalignment. These should be reported as well.⁵⁸ They tell regulators what models are *willing to do now*, even if they get caught by safeguards, or if they only take these actions in evaluations.⁵⁹

Reportable behaviours are different from near misses. A near miss is reportable only if there was a credible path to serious harm. A listed behaviour must be reported whatever the context. For example, a model that tries to deceive its developer during a sandboxed evaluation would probably not count as a near miss, but regulators should still know it happened, so they understand what models are willing to do.

⁵³ Kara, Z. (2026). [Improving Frontier AI Incident Reporting Regimes](#). Centre for the Governance of AI.

⁵⁴ Kara, Z. (2026). [Improving Frontier AI Incident Reporting Regimes](#). Centre for the Governance of AI.

⁵⁵ Kara, Z. (2026). [Improving Frontier AI Incident Reporting Regimes](#). Centre for the Governance of AI.

⁵⁶ Commonwealth of Australia. (2021). [Transport Safety Investigation Regulations 2021](#).

⁵⁷ Pho, B. (2026). [California's first AI-safety law didn't cover the first rogue AI hacks](#). Mission Local.

⁵⁸ Kara, Z. (2026). [Improving Frontier AI Incident Reporting Regimes](#). Centre for the Governance of AI.

⁵⁹ Kara, Z. (2026). [Improving Frontier AI Incident Reporting Regimes](#). Centre for the Governance of AI.

A developer should only have to report an incident once, even if it qualifies under more than one trigger. Because the scheme would cover only the small number of companies training *frontier models*, lower thresholds will not overwhelm the regulator.

Capture model theft or modifications

Unauthorised theft of frontier model weights is a serious risk. Theft of model weights gives adversaries, including hostile states, capabilities they may not be able to build themselves. They also let anyone who holds them strip out safeguards built into the model, enabling dangerous use. Unauthorised modification is also dangerous. Tampering with model weights can remove safety training or introduce hidden behaviours in a model that is then used as normal.⁶⁰

Neither California nor the EU incident reporting obligations fully capture model theft and alteration.

- California covers unauthorised access to, modification of, or exfiltration of model weights, but only *where this results in death or bodily injury*.⁶¹ But regulators need to know when theft happens, not when it causes harm later.
- The EU captures theft only through its voluntary Code of Practice. Companies can rely on the Code to demonstrate compliance with the *AI Act*, but they do not need to. Not every frontier developer has signed the Code. It is unclear whether the Code applies to models that are not released, and it does not cover unauthorised modification of model weights.⁶²

Australia should adopt California's description of "unauthorised access to, modification of, or exfiltration of model weights" as a reportable event, but remove the additional harm requirement.

Companies should be required to report information about AI capabilities

Transparency reporting tells regulators what frontier AI can do, what risks it poses, and how frontier AI companies are managing those risks. Other jurisdictions have begun to require this. The *EU AI Act* and California's *Transparency in Frontier Artificial Intelligence Act* require developers of the most capable models to describe their models and intended uses, assess dangerous capabilities, report the results, and explain how they manage risks. The EU also requires developers to report the compute used to train the model. Australia should adopt these requirements at minimum.

However, international obligations fall short in several areas. Closing these gaps would make Australia a world leader in frontier AI transparency.

Cover internal models

Frontier AI companies use their most capable models internally for weeks or months before public release, and some models may never be released. For similar reasons described in the *Incident Reporting* subsection, having information about internal models is important. Further, understanding internal models is important to understanding capability development speeds. Anthropic has said it is "plausible that models equipped with highly effective scaffolding may not be very far away" from the ability to "fully automate the work of an entry-level, remote-only Researcher at Anthropic".⁶³ If companies succeed in using models to automate AI research, capabilities may improve rapidly. Regulators should be warned if so.

⁶⁰ Delaney, O. et al. (2026). [Risk reporting for developers' internal AI model use](#). Institute for AI Policy and Strategy.

⁶¹ [Transparency in Frontier Artificial Intelligence Act 2025](#) (California), s 22757.11(d)(1).

⁶² [Regulation \(EU\) 2024/1689 \(Artificial Intelligence Act\)](#) (European Union), arts 2(8), 3(3), 3(49), 55(1)(c)–(d), 55(2); European Commission. (2025). [Code of Practice for General-Purpose AI Models: Safety and Security Chapter](#), Commitments 6 and 9, Measure 9.3, Glossary.

⁶³ Anthropic. (2025). [System card: Claude Opus 4.5](#).

The *EU AI Act* likely does not cover internal models.⁶⁴ The Californian Act does, but doesn't require enough detail from AI companies.⁶⁵ Reports on internal models should be detailed, covering:⁶⁶

- How internal models compare to public ones, including how they score on benchmarks and where their capabilities exceed public models in dangerous domains such as cyber offence.
- Safeguard measures, how those measures have been stress-tested, and results on misalignment and monitorability evaluations.
- Access and use, including who can access internal models and what for, and prohibited uses.
- Which concerning behaviours the developer monitors for, and how it responds when it finds them.
- How the internal model contributes to AI research and development. This is discussed further below.

Include detailed research and development metrics

Frontier AI companies are trying to automate AI research. If they succeed, capabilities may advance rapidly. Regulators should know the degree to which frontier models are helping AI companies build future models.

Today, information about the use of AI to automate AI research depends almost entirely on patchy and voluntary disclosures. The EU AI Act's Code of Practice asks developers to assess whether a model can automate AI research when assessing whether an AI poses systematic risk, and to describe at a high level how they use models to develop new ones.⁶⁷ But this information is not granular, may not apply to internal models, and not every frontier company has signed the code.

Regulators should work with the AISI and frontier AI companies to understand the best information to report. For example, Australia may require captured entities to report.⁶⁸

- The share of compute used for internal AI research agents, compared with training and external deployment.
- Compute spending as a share of total research and development spending.
- The extent of AI involvement in high-stakes decisions, such as launching training runs or deploying models.
- Lists of the actions AI systems may take without human approval.
- Incidents where AI systems attempted to sabotage research or hide misbehaviour.
- Staff-reported productivity gains.

Regulators should update rules and metrics as better ones emerge.

Include detailed biosecurity capability evaluations for all models

AI could help people create biological weapons. Frontier models now outperform human experts on some biosecurity-relevant tasks, and exceed PhD-level virologists at troubleshooting lab protocols.⁶⁹ Regulators need to know how much an AI model uplifts someone trying to build a biological weapon. Organisations that run these evaluations for developers say they 'strongly support greater transparency and accountability in AI deployment'.⁷⁰

⁶⁴ For discussion on the legal detail here, see Pistillo, M. (2026). [Internal deployment in the AI Act](#). Cambridge Commentary on EU General-Purpose AI Law.

⁶⁵ See, eg, Charnock, J. et al. (2026). [What should frontier AI developers disclose about internal deployments?](#). arXiv.

⁶⁶ This list is adapted from the research of Charnock, J. et al. (2026). [What should frontier AI developers disclose about internal deployments?](#). arXiv.

⁶⁷ European Commission. (2025). [Code of Practice for General-Purpose AI Models: Safety and Security Chapter](#).

⁶⁸ This list is adapted from the research of Chan, A., Padarath, R., Kwon, J., Greaves, H., & Anderljung, M. (2026). [Measuring AI R&D Automation](#), Centre for the Governance of AI.

⁶⁹ Bengio, Y., et al. (2026). [International AI Safety Report 2026](#); Anthropic. (2026). [System card: Claude Fable 5 & Claude Mythos 5](#). Götting, J., et al. (2025). [Virology Capabilities Test \(VCT\): A multimodal virology Q&A benchmark](#). arXiv.

⁷⁰ SecureBio. (2025). [SecureBio's AI Team: An Overview of Our Biorisk Evaluations](#).

Different international approaches require different levels of biosecurity reporting:

	EU AI Act	California TFAIA
Public Models	Through the <i>EU Code of Practice</i> , developers must evaluate biological risk as part of the detailed model report before the model is placed on the market, including the evaluation methods, tests, and results behind the assessment. ⁷¹	Large developers must publish summaries of their catastrophic risk assessments, including for biological weapons, and the results. Developers choose which assessments to run. They don't have to disclose the underlying evaluation methods or tests. They can redact information to protect trade secrets, cybersecurity, public safety or national security. ⁷²
Internal Models	Coverage of internal models remains legally uncertain. Research and development exemptions may leave many internal deployments outside scope. ⁷³	Large developers must confidentially give the state emergency services agency a summary of any catastrophic risk assessments from internal use, which can include biological weapons risk. The summaries are not published, and the law does not expressly require results or evaluation methods. ⁷⁴

Australia should combine the EU's level of detail and transparency about the underlying evaluations with California's coverage of internal models. Frontier AI companies should give the regulator full, state-of-the-art biosecurity evaluation results for every frontier model, whether deployed or used only internally.

Including information already disclosed to overseas regulators

Frontier AI companies give overseas regulators information about their most capable models. Over time, more countries will introduce requirements and update existing international regulations. Australia should take advantage of this. Although we can update our own regulations, we should also have access to new information as it is reported to our international counterparts.

Australia should require frontier AI companies to provide us with the reports they have provided to international regulators. This would cost companies very little, but it would help ensure Australia is never less informed than our international counterparts.

Australian regulators already require companies to communicate overseas safety developments. For example, the TGA requires companies to share the risk management plans they give overseas regulators. When a company applies to register a new medicine, it submits the most recent plan it gave the European regulator, with an Australia-specific annex that sets out any differences for Australia.⁷⁵ Further, companies must report when a comparable overseas regulator takes a major safety action, like withdrawing their product from the market or adding a warning. The TGA expects companies to communicate this information in 'no more than 3 calendar days from global awareness'.⁷⁶

⁷¹ [Regulation \(EU\) 2024/1689 \(Artificial Intelligence Act\)](#) (European Union), art 55(1)(a); European Commission. (2025). [Code of Practice for General-Purpose AI Models: Safety and Security Chapter](#), Commitment 7, Measures 7.3 and 10.2, Appendix 1.4.

⁷² [Transparency in Frontier Artificial Intelligence Act 2025](#) (California), ss 22757.11(c)(1)(A), 22757.12(a)(2), (c)(2), (f).

⁷³ For detailed discussion, see, Pistillo, M. (2026). [Internal deployment in the AI Act](#). Cambridge Commentary on EU General-Purpose AI Law.

⁷⁴ [Transparency in Frontier Artificial Intelligence Act 2025](#) (California), ss 22757.12(a)(4), (a)(10), (d), 22757.13(b), (f).

⁷⁵ Therapeutic Goods Administration. (2019). [Risk management plans for medicines and biologicals: Australian requirements and recommendations](#).

⁷⁶ Therapeutic Goods Administration. (2023). [Pharmacovigilance responsibilities of medicine sponsors: Australian recommendations and requirements](#).

Risk Management Obligations

Australia's crisis management and critical infrastructure laws

Australia has laws and policies that protect Australians in a crisis. Expanded domestic AI infrastructure would bring new and valuable stakeholders to the table. AI may create new types of crises, amplify existing crises, and help respond to and recover from other crises. Government should ensure existing laws and policies are ready to take advantage of new stakeholders and infrastructure.

AI in the Australian Government Crisis Management Framework

The National AI Plan commits the Government to considering how AI-related harms are managed under the Australian Government Crisis Management Framework (AGCMF).⁷⁷

Under the AGCMF, government works with industry through mechanisms like the National Coordination Mechanism. AI companies with a significant presence in Australia are more likely to take part in such mechanisms because they have local staff, physical assets, and an ongoing relationship with government.

Good Ancestors has argued that AI incidents, which could be much broader than cyber incidents as AI capability grows in many domains, should become an explicit hazard in the AGCMF, with a designated lead agency and a dedicated AI crisis plan.⁷⁸

An AI crisis plan could do three things:

1. ensure AI companies, data centre operators, relevant civil society groups, and the Australian AI Safety Institute know they have a role to play during emergencies, including through the National Coordination Mechanism and forums like the Trusted Information Sharing Network (TISN)
2. set out practical response options, such as restricting model access or rolling back an AI service, if they cause a crisis via a pre-agreed authorising environment, and
3. be tested through the National Emergency Management Agency national exercising program. Companies hosting AI infrastructure in Australia would be natural early participants.

The TISN could play a globally important role in AI safety. The TISN is designed to allow members of the critical infrastructure community to collaborate to enhance the resilience of their organisations and sector. This kind of collaboration is globally important and the frontier of AI, and many other countries do not have similar paths to facilitate the lawful sharing of threat information.

Combined, these measures would help ensure Australia does not face its first AI crisis with untested plans and unfamiliar partners.

AI in the Security of Critical Infrastructure Act

While the SOCI Act was not built for crisis management, it is one of the government's key tools when a crisis affects critical infrastructure. The *SOCI Act* is being reformed to better cover AI infrastructure. In July 2026, Good Ancestors recommended:⁷⁹

- keep AI inference services within service-based capture,
- capture entities that do AI training, not just the data centre operators, and

⁷⁷ Department of Industry, Science and Resources. (2025). [National AI Plan](#). Australian Government.

⁷⁸ Grundy, E. & Sadler, G. (2026). [China has an AI crisis plan. Australia should, too](#). The Strategist.

⁷⁹ Good Ancestors, (2026). *Capturing AI Training and Inference Under the SOCI Act*.

- ensure at least one pathway reaches AI service providers that other critical infrastructure depends on.

These changes would ensure government can reach the AI entities best placed to respond when an AI-created or -mediated crisis hits.

Compute verification obligations

Compute verification is a set of tools and techniques for checking what is happening inside an AI data centre. This includes whether the facility is training a new model or running existing ones, whether it is running the model its operator claims, and whether it is complying with agreed rules. Tools can operate without exposing sensitive information (e.g., model weights or user data).

Why is compute verification important?

Compute verification matters because many of the obligations in this submission rely on information only companies hold. In the near term, verification supports oversight of AI infrastructure on our soil. If foreign companies train or run frontier models in Australian data centres, verification lets government independently confirm what those facilities are doing.⁸⁰

In the longer term, verification could support international agreement on how AI risks are managed. Nuclear arms control agreements relied on analogous verification technology for decades.⁸¹ Verification technology was essential to international agreements being viable. Over the next few years, there may be a stronger push for slowing or pausing frontier development due to catastrophic risk concerns or ahead of recursive self-improvement leading to artificial superintelligence.⁸² Agreements may depend on parties verifying one another's compliance.

The science of verification is nascent,⁸³ but researchers think recording and checking what happens inside data centres could be possible within a few years.⁸⁴ The Australian AI Safety Institute could lead this effort. The work draws on skills Australia already has, such as engineering, cybersecurity, cryptography, national security, and diplomacy.

What might compute verification obligations look like?

In the short term, frontier AI companies that train in Australia could be required to work with AISI researchers and government policymakers on compute verification.⁸⁵ That could mean letting researchers into their facilities, testing early verification tools, and lending engineers to AISI on compute verification projects. At least one frontier developer has publicly committed to helping build these tools.⁸⁶

In the long term, frontier AI companies should agree to install verification tools once government is confident they work. These could be devices connected to a data centre's network or power supply, software running on its servers, or features built into the AI chips themselves.

⁸⁰ The Future Society. (2026). [How to make international AI verification a reality](#).

⁸¹ Favaro, M. & Clark, J. (2026). [When AI builds itself](#). Anthropic Institute.

⁸² The Future Society. (2026). [How to make international AI verification a reality](#).

⁸³ See further discussion in Baker, M. et al. (2025). [Verifying international agreements on AI: Six layers of verification for rules on large-scale AI development and deployment](#). RAND Corporation.

⁸⁴ Baker, M. et al. (2025). [Verifying international agreements on AI: Six layers of verification for rules on large-scale AI development and deployment](#). RAND Corporation

⁸⁵ Experts have called for computer verification to be an area of specialisation within the AISI: Good Ancestors. (2026). [Designing Australia's AI Safety Institute: Expert survey report](#).

⁸⁶ Favaro, M. & Clark, J. (2026). [When AI builds itself](#). Anthropic Institute.

Embedded safety evaluators

“Embedded evaluators” involves an AI company committing to ongoing, “employee like” access to a team of embedded third party evaluators whose job is to verify adherence to commitments, safety practices, report incidents, and help assess the alignment of models.⁸⁷

Embedded safety evaluators may help to ensure transparency and appropriate incident reporting obligations where regulators are unsure of what is required. Embedded evaluators are better positioned than external regulators to assess risks emerging from fast-moving internal systems and practices.⁸⁸ Evaluators with hands-on access are better able to complete in-depth assessments, identify risks and remediate vulnerabilities before they become incidents.⁸⁹ Functions of third party evaluations include confirming internal work, strengthening the robustness of assessments, and supplementing expertise and resources.⁹⁰

Anthropic has unilaterally committed to implementing embedded evaluators and called on governments to require other frontier AI companies to do the same.⁹¹ Sam Altman, CEO of OpenAI, agreed claiming that OpenAI would also commit to having independent evaluators with “employee like” access.⁹²

- Demis Hassabis, Google DeepMind Chairman, broadly agreed with Amodei’s proposal. He proposed a Frontier AI Standards Body with a board of independent technical experts and open-source representatives. This body would act as an independent regulator, encouraging frontier AI companies to voluntarily share their models before release for evaluation.⁹³
- Elon Musk, xAI CEO, in 2023 called for a “referee” to ensure companies take actions that are safe and in the best interests of the public.⁹⁴

Australia should take AI leaders at their word and impose on them the regulations that they’re asking for.

What might embedded safety evaluators look like?

In a pilot embedded assessment, AI safety organisation METR found flaws in Anthropic’s live attribution logic.⁹⁵ METR described this as a “highly non-obvious” flaw without specific knowledge of Anthropic’s internal infrastructure.

Current internal monitoring and security systems have demonstrated weaknesses. The OpenAI-Hugging Face incident is one example where internal monitoring failed to detect an incident in a timely manner. Anthropic has similarly reported cyber incidents that initially went undetected.⁹⁶ Embedded assessments may help reduce the frequency and severity of incidents.

For Australia, this would include requiring AI companies to maintain independent embedded evaluators designed to assess a range of risk areas. This auditing should be conducted by third parties with no direct financial stake in the company, which is standard practice in other industries.⁹⁷ Evaluators would require

⁸⁷ Amodei, D. (2026). [We Must Pace the Frontier](#).

⁸⁸ Charnock et al. (2026). [Embedded Assessments for Frontier AI](#). ArXiv.

⁸⁹ Charnock et al. (2026). [Embedded Assessments for Frontier AI](#). ArXiv.

⁹⁰ Frontier Model Forum. (2025). [Third-Party Assessments](#).

⁹¹ Amodei, D. (2026). [We Must Pace the Frontier](#).

⁹² Altman, S. (2026). [X \(Twitter\)](#).

⁹³ Hassabis, D. (2026). [X \(Twitter\)](#).

⁹⁴ Shepardson, D., Warburton, M., and Stone, M. (2023). [Tech Titans Meet US Lawmakers, Musk Seeks ‘Referee’ for AI](#). Reuters.

⁹⁵ Charnock et al. (2026). [Embedded Assessments for Frontier AI](#). ArXiv.

⁹⁶ Charnock et al. (2026). [Embedded Assessments for Frontier AI](#). ArXiv.

⁹⁷ Brundage et al. (2026). [Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies](#). ArXiv.

access to technical and organisational infrastructure, provided through secure evaluation environments if required for addressing privacy or security concerns.⁹⁸

The pilot conducted by METR suggests that internal agent monitoring is a valuable focus for embedded assessments, along with assessing the effectiveness of sandboxes and other security controls, and examining whether models stay aligned throughout their lifecycle.⁹⁹ Evaluations should be standardised and apply predefined criteria describing sufficient evidence of safety, security, and risk management.¹⁰⁰

Conclusion

AI development is rapid, accelerating, and on track to upend geopolitics and the global economy. Australia will be left behind if we fail to secure many valuable positions in a world with highly capable AI. We must then use those positions to secure the national interest.

We have provided a number of possible conditions. Government should give the most weight to conditions Australia cannot obtain any other way. Frontier model access is the most important. As the most capable models become restricted, linking access to permission to train may be the only reliable way for Australia to secure access.

Other conditions deserve serious consideration. Each would bring real benefits to Australia, and each carries different costs for frontier AI companies. Government should weigh them against each other, our ability to secure them later or on the market, and against the risk of frontier AI companies choosing to train elsewhere.

⁹⁸ Brundage et al. (2026). [Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies](#). ArXiv.

⁹⁹ Charnock et al. (2026). [Embedded Assessments for Frontier AI](#). ArXiv.

¹⁰⁰ Brundage et al. (2026). [Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies](#). ArXiv.