Good
Ancestors
Policy

AI Safety as Economic Opportunity

2025-26 Pre-Budget Submission Treasury

January 2025

Good Ancestors is an Australian charity dedicated to improving the long-term future of humanity. We care about today's Australians and future generations. We believe that Australians and our leaders want to take meaningful action to combat the big challenges Australia and the world are facing. We want to help by making forward-looking policy recommendations that are rigorous, evidence-based, practical, and impactful.

Good Ancestors has been engaged in the AI policy conversation since our creation, working with experts in Australia and around the world while connecting directly with the Australian community.

Good Ancestors is proud to help coordinate Australians for AI Safety.

Our thanks go to the volunteers who provided input to this submission and who care so passionately about being good ancestors to future generations of Australians.

Executive Summary

AI will soon be a significant economic force. But Australians are more worried about AI than any other nation. Government's substantial investments in AI adoption[1] are hamstrung by a lack of trust. Efforts to build trust are missing the mark.

Government should fund the creation of an Australian AI Safety Institute (AISI). This has three key benefits:

1. An AISI is a tangible action to demonstrate that Government is taking serious action, alongside global leaders, to address safety concerns.
2. An AISI would catalyse Australia's growing AI Assurance Technology (AIAT) industry. The global AIAT market will reach $276b by 2030, and Australia is well placed to act now to secure a significant share.
3. Australia has already committed to creating an AISI, and has made other commitments that an AISI would fulfil.

Analysis of AI's potential contribution to the Australian economy shows fast AI adoption being worth $70b per year more by 2030 than slow AI adoption.[2] Even if creating an Australian AISI has only a moderate impact, each dollar spent could return thousands of dollars in economic activity through improved trust alone. This would be in addition to the benefits of fostering the domestic AIAT industry and reducing the downside risks that AI presents.

[1] 'Developing a National AI Capability Plan'. Department of Industry Science and Resources. 13 December 2024.

[2] See discussion of downstream economic activity, page 9.

Table of Contents

AI will be a major economic force

Artificial intelligence (AI) has evolved from a potential digital innovation to the **next frontier of economic transformation**. Leading governments and companies recognise AI as key to **national competitiveness, societal well-being, and strategic autonomy**.

There are two aspects to the AI economy:

1. **Upstream economic activity**: The development of AI models and AI systems, including inputs to those systems and various kinds of AI products.
2. **Downstream economic activity:** The productive use of AI tools to enhance existing business models or create new ones.

We can have a high degree of confidence that AI will be a major economic force because businesses, the market and leading governments are backing it. The fundamentals also make sense – AI can already perform at human level on many tasks, meaning that many tasks will shortly be able to be performed to a high standard at a fraction of the cost. PwC calls AI a "$15.7 trillion game changer",[3] and the IMF says:[4]

> *[A]lmost 40 percent of global employment is exposed to AI. Historically, automation and information technology have tended to affect routine tasks, but one of the things that sets AI apart is its ability to impact high-skilled jobs. As a result, advanced economies face greater risks from AI—but also more opportunities to leverage its benefits.*

Businesses and the market are backing AI

Capital is flowing into AI. In the fourth quarter of 2024, over half (50.8%) of global venture capital funding went to AI companies.[5] **Capital spending on AI rivals the mainframe era of the late 1960s and the fiber optic deployment of the late 1990s**.[6] Demand for AI data centres is **rising 167% year-on-year**—far exceeding previous tech booms.[7] **Generative AI Funding** jumped to **USD 25.2b** in 2023, representing an eightfold increase from 2022, pointing to surging confidence in AI technologies.[8]

McKinsey reports that, as of early 2024, 72% of surveyed businesses had adopted AI, up from 50% in 2023.[9] Although methodologies differ, AI adoption in comparable countries greatly exceeds that of Australia.[10] Organisations using AI have reported an average of

[3] In USD from 'Global Artificial Intelligence Study'. PWC

[4] 'AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity.' International Monetary Fund.

[5] 'AI dominates venture capital funding in 2024'. fDi Intelligence.

[6] 'A severe case of COVIDIA: prognosis for an AI-driven US equity market'. JP Morgan Private Bank U.S.

[7] 'Infrastructure Is Destiny: Economic Returns on US Investment in Democratic AI', OpenAI, September 2024.

[8] 'Artificial Intelligence Index Report 2024. Chapter 4: Economy'. The AI Index. Stanford University.

[9] 'The state of AI in early 2024'. McKinsey. 30 May 2024.

[10] 'Exploring AI adoption in Australian businesses'. Department of Industry Science and Resources. 17 December 2024.

42% cost reductions and **59% revenue gains**.[11] The figures are only likely to grow as AI capability improves.

Deepseek's disruption of markets in January 2025 demonstrated the ability of a new firm to make a capable model with only simple improvements to algorithms.[12] This suggests that significant further gains in AI capability remain readily available. Deepseek's low-cost model also demonstrates that Australia still has a window of opportunity to become an innovator.

Major countries are also backing AI

Leading governments are also investing in AI. The United States invested **USD 67.2 b** in AI infrastructure in 2023. Executive Orders under the Biden administration aimed to accelerate AI data-centre expansion on federal land and to maintain American leadership in safe AI innovation.[13] The Biden administration's October 2024 **National Security AI Framework** secured AI supply chains, signalling the strategic implications of AI.[14] The Trump administration has doubled down on AI, announcing a **USD 500b Stargate Project**—a collaboration between Arm, Microsoft, NVIDIA, Oracle, and OpenAI.[15]

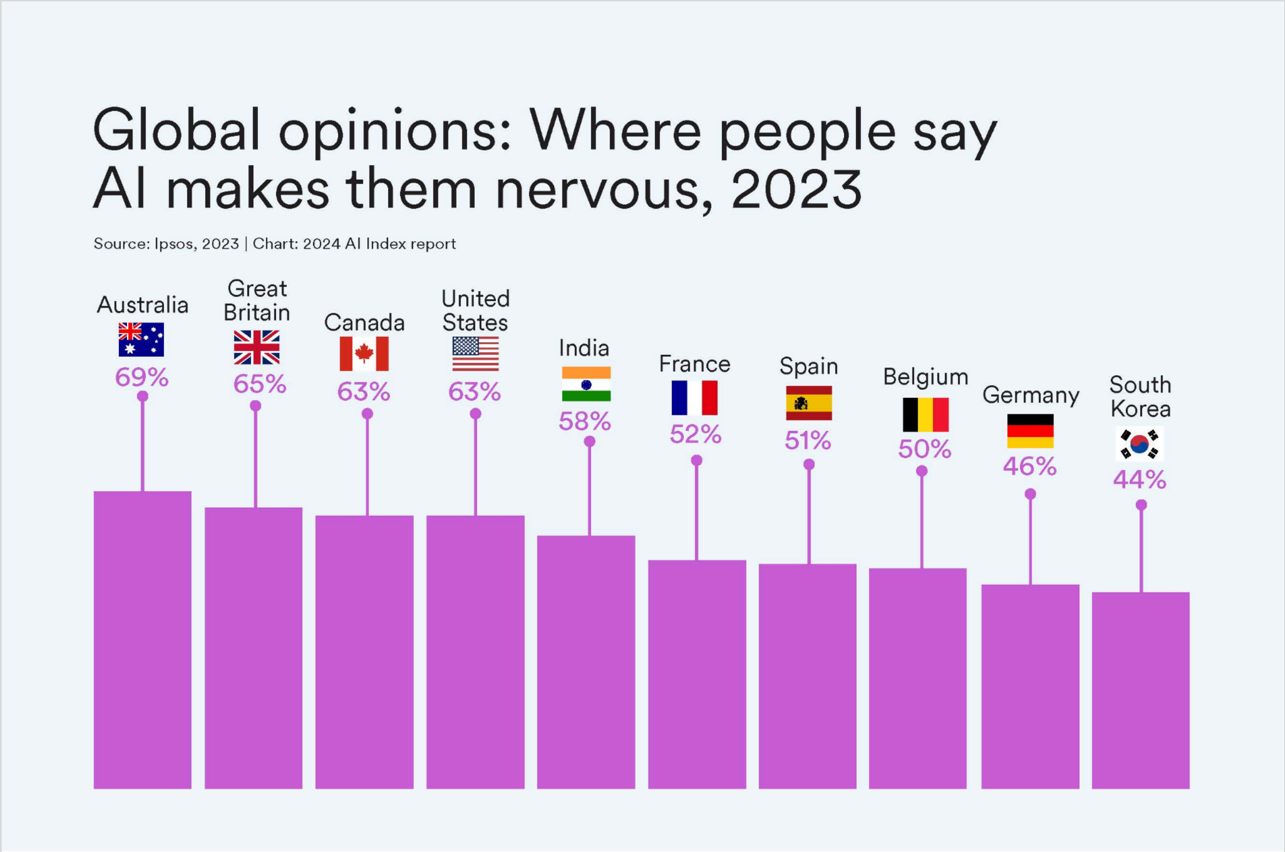
China spent **USD 7.8 b** on core AI infrastructure in 2023, underpinned by the **New Generation Artificial Intelligence Development Plan**, which aims for **global AI leadership by 2030**. China's **September 2024 AI Safety Governance Framework** emphasises balancing innovation with security, revealing a strong government mandate to dominate AI while controlling risks.[16]

These trends are global. The **UAE's $1.5 trillion sovereign wealth investment portfolio treats AI as a national priority**. This includes developing the Falcon large language model and targeting $100 billion in assets under management for its AI investment firm, MGX. Strategic collaborations, such as Microsoft's $1.5b investment in G42, further solidify the UAE's role as a global AI innovation hub.

[11] 'Cost decrease and revenue increased from AI adoption by function.' McKinsey & Company Survey 2023.
[12] China's DeepSeek suggests AI can be done cheap. The fallout could be costly - ABC News. ABC News. January 2025.
[13] 'Executive Order 14141 on Advancing United States Leadership in Artificial Intelligence Infrastructure'. The White House. 14 January, 2025.
[14] 'New rules for US national security agencies balance AI's promise with need to protect against risks'. AP News.
[15] 'Announcing The Stargate Project'. OpenAI. Note: The extent of Government funding is currently unclear.
[16] 'China Unveils Comprehensive AI Security Governance Framework at Cybersecurity Forum'. BABL AI. 13 September 2024.

Australians uniquely mistrust AI

Australians uniquely mistrust AI—being more concerned by its risks than any other country for which there's data.[17] The Department of Industry, Science and Resources correctly diagnoses **public mistrust** as a prime factor restricting AI adoption—stressing the importance of **credible oversight** in building societal confidence. The Office of the Australian Information Commissioner also shares this view.[18]



Global opinions: Where people say AI makes them nervous, 2023

Source: Ipsos, 2023 | Chart: 2024 AI Index report

*Figure 1: **69%** of Australians express apprehension about AI, ranking us at the top of global anxiety surveys.*[19]

Drilling into the details, a technical report cited by the Department of Industry shows that Australians' concerns relate primarily to the "big risks" that highly capable AI could pose.[20] **Australians are most concerned about catastrophic risks and AI safety**, rather than its responsibility or ethics.

[17] Australians most nervous globally about AI'. Ipsos. 11 July 2023.

[18] OAIC submission to the Department of Industry, Science and Resources – Safe and responsible AI in Australia discussion paper. Office of the Australian Information Commissioner. 21 September 2023.

[19] Australians most nervous globally about AI'. Ipsos. 11 July 2023.

[20] 'Survey Assessing Risks from Artificial Intelligence: Technical Report'. Ready Research, University of Queensland. 2024.

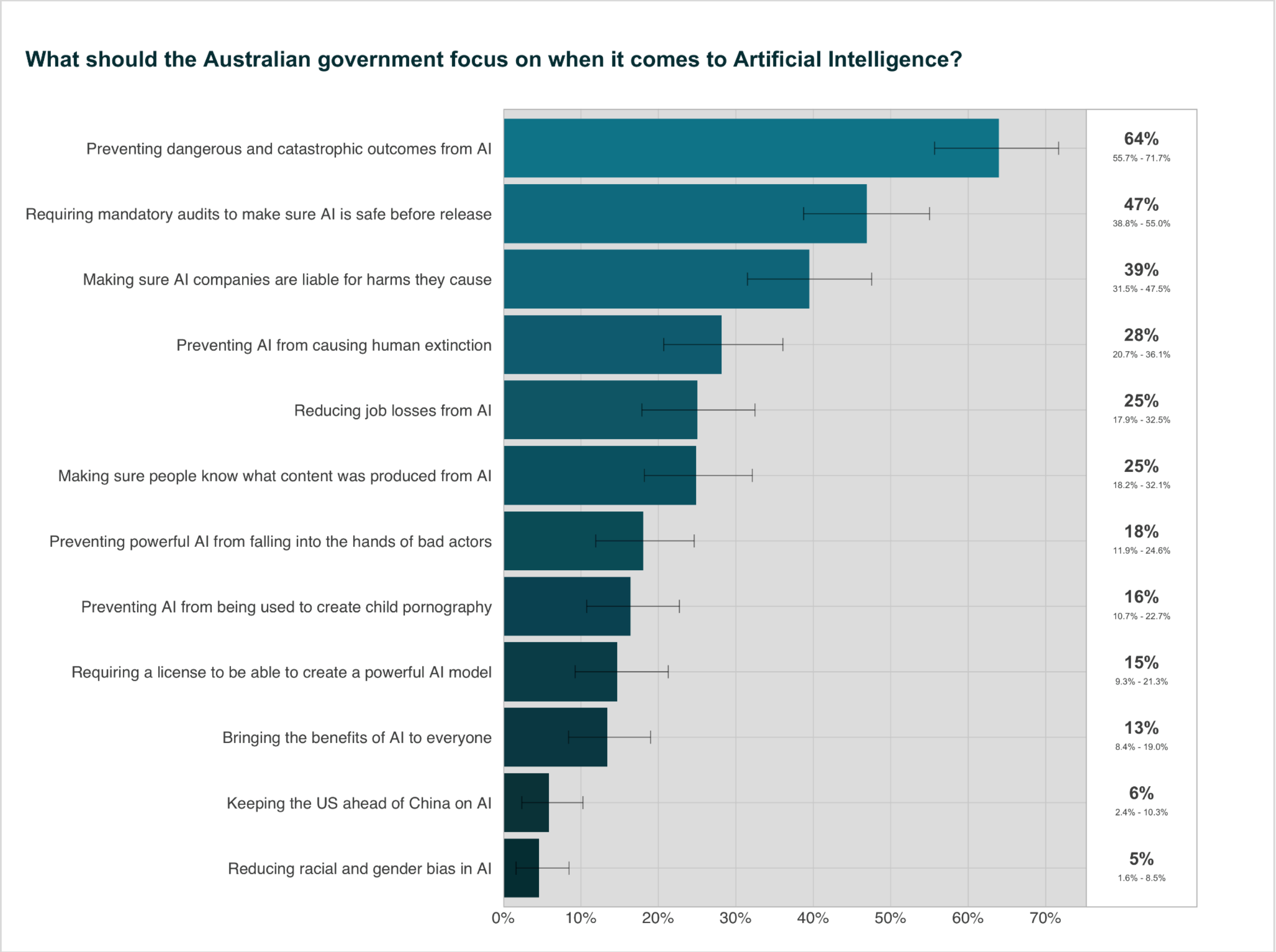


Figure 2: Research shows that Australians' top priority for government focus is AI safety, specifically risks like preventing catastrophic outcomes. Broader ethical issues are a lower priority.

Government has not provided specific funding for AI safety despite the significance of the risk and the level of public interest. Government investments have focused on the National AI Centre, whose mission is to "support and accelerate Australia's AI industry"[21] and on expanding the University of Adelaide's Australian Institute for Machine Learning to "support businesses to develop AI-enabled products and services, automate processes and improve productivity."[22]

Although these initiatives engage in some work relating to ethical AI, they are not designed to address Australia's core concerns and instead mostly focus on accelerating AI capability and driving adoption. KPMG's 2020 survey of Australians shows that the industry's relationship with AI is the cause of the trust problem, not the solution.[23] If anything, **these investments have reinforced the public perception that Government is only paying lip service to public concern and is superficially addressing risks**.

[21] 'National Artificial Intelligence Centre'. Department of Industry Science and Resources. Accessed 22 January 2025.
[22] 'Responsible AI focus for new research centre'. Newsroom. University of Adelaide. 9 December 2024.
[23] 'Trust in Artificial Intelligence: Australian insights 2020'. KPMG and University of Queensland. October 2020.

Upstream economic activity: the AI Assurance Technology industry

Australia is not positioned to match the infrastructure spending of superpowers,[24] but we are able to find our niche. One option is the AI Assurance Technology industry (AIAT).

The generative and potentially autonomous nature of AI creates unique challenges for reliably using AI to achieve business goals as well as ensuring compliance with existing general legal frameworks and emerging AI-specific regulations, like those in force in the EU or proposed in Canada and Australia.

> *“Ubiquitous artificial intelligence is seemingly upon us—maybe it’s already here. Yet we do not know how to effectively control, align or guarantee trustworthiness in these systems. In addition to sensible regulations, we will need to scale technological countermeasures to create the beautiful future that this technology can enable.”*
>
> *- Sara Rywe, Partner, byFounders*[25]

The AIAT industry will grow to address this gap. **AIAT is the software, hardware, and services that enable organisations to more effectively, more efficiently, or more precisely mitigate the risks of AI.** The AIAT industry services both upstream AI developers and downstream AI adopters.

The global AIAT industry was worth USD 1.6b in 2023[26] and is forecast to grow rapidly to USD 276b by 2030. If predictions about the exponential growth of AI capability and the AI industry are correct, AIAT could be worth trillions.

Leveraging proven expertise in safety and assurance

Australia’s success in globally competitive safety-critical industries—such as mining safety, aviation standards, and food biosecurity—demonstrates our capability to develop and export robust assurance frameworks. This provides a strong foundation for establishing Australia’s leadership in AIAT, ensuring that Australian solutions are trusted and adopted worldwide.

The AIAT market is a particularly good fit for Australia. AIAT is an “upstream” economic activity that would position **Australia as a direct contributor to AI products** that will be

[24] Good Ancestors would support a broad change in policy where Government is overtly ambitious on AI and “bets big” to shape our future. This could involve significant investment in compute infrastructure and frontier models. Recent news around Deepseek suggests that lower-cost innovation is still possible. This paper assumes that there will not be a substantial policy change on this topic.
[25] ‘Risk & Reward - 2024 AI Assurance Technology Market Report’. Juniper Ventures. Accessed 22 January 2025.
[26] ‘Risk & Reward - 2024 AI Assurance Technology Market Report’. Juniper Ventures. Accessed 22 January 2025.

consumed globally. Further, AIAT is less constrained by the need for billion-dollar infrastructure investments required for direct involvement in training frontier models.

The AIAT market is also a good fit for Australia's national strengths, including socio-technical research skills, our regional position, and the lifestyle factors that may attract global talent to Australia. Equally, without intervention, Australia will continue to lose talent and fall further behind projections.

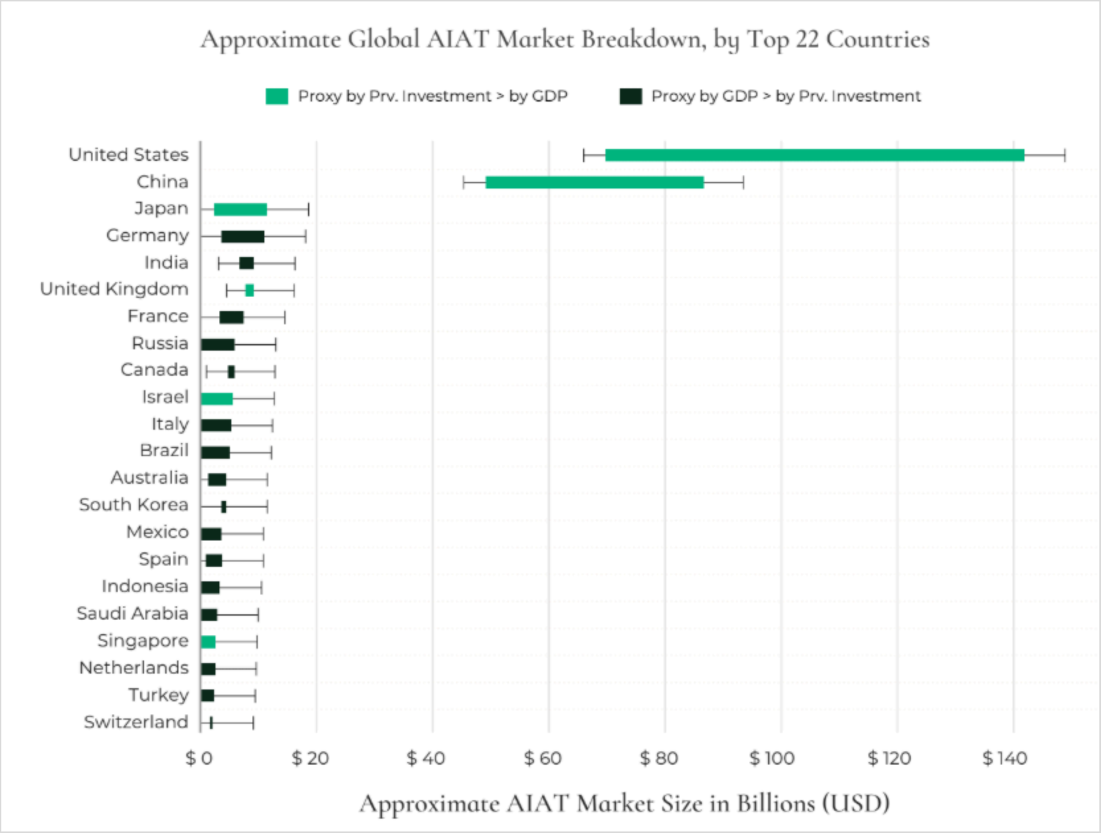


Figure 3: With appropriate investment, Australia could be a global leader in the AIAT market.[27]

AI safety is seen by other nations as a high-return strategy. The UK has developed an AIAT industry plan, stating that "a proactive, targeted programme of work will help to realise this potential, accelerating innovation and investment in AI assurance".[28] The **UK (GBP 100m)** and **Canada (CAD 50m)** are also investing heavily in **dedicated AI safety** initiatives, recognising them as **high-return** pathways to global competitiveness.

By following suit, Australia can **differentiate** itself from infrastructure races, forge **international partnerships**, and **shape global AI norms**. Australia already has promising AIAT startups, but they need support to become global players.

[27] 'Risk & Reward - 2024 AI Assurance Technology Market Report'. Juniper Ventures. Accessed 22 Jan 2025.
[28] Assuring a Responsible Future for AI - GOV.UK. UK Department for Science. Accessed 29 January 2025.

Downstream economic activity: Only if trust is fostered

AI's potential to contribute to Australia's economy is widely recognised. A range of organisations using different methodologies have tried to quantify the annual economic value by 2030, including:

- $45-115b[29] by Tech Council of Australia[30]
- $280b by Google[31]
- $170-$600b by McKinsey,[32] and
- $200b by Kingston AI Group.[33]

While the methodology of estimates varies, AI is on track to be a significant economic factor, perhaps approaching the scale of the mining industry.[34]

These reports were all published before the announcement of "**Priority Access" via America's new Tier I Export Controls.**[35] **Under these new export controls, Australia is one of a small number of nations granted unrestricted access** to advanced AI chips, offering a **strategic advantage** for local AI development. This access raises the stakes for innovation but also increases risks if safety and security are not managed.

Each report highlights that Australians distrust AI and that distrust will be a key hurdle to adoption. While the reports mention trust, they typically do not explain whether Australia's uniquely low trust factors into their assessment. As set out above, Australians are worried about the risks of AI, with a focus on catastrophic risk, and low trust means slow adoption.

The Tech Council of Australia's analysis does explore the significance of adoption pace to the quantity of economic benefit—its modelling shows that faster adoption could generate up to $115 billion in annual value by 2030, compared to just $45 billion in its slow adoption scenario. This 156% difference in value is driven primarily by adoption rates, and hence trust.

[29] $45b is a slow adoption scenario, $115b is a fast adoption scenario. See further discussion, below.
[30] 'Australia's Generative AI opportunity'. Tech Council of Australia and Microsoft. July 2023.
[31] 'Economic Impact Report: Turning Australia's AI opportunities into impact with Google'. Access Partnership. June 2024.
[32] 'Australia's automation opportunity: Reigniting productivity and inclusive income growth'. McKinsey & Company. 3 March 2019.
[33] 'Australia's AI Imperative'. Kingston AI Group. 16 April 2024.
[34] 'Australia: value added by mining industry 2023'. Statista. 24 June 2024.
[35] 'With Its Latest Rule, the U.S. Tries to Govern AI's Global Spread'. Carnegie Endowment for International Peace. 13 January 2025.

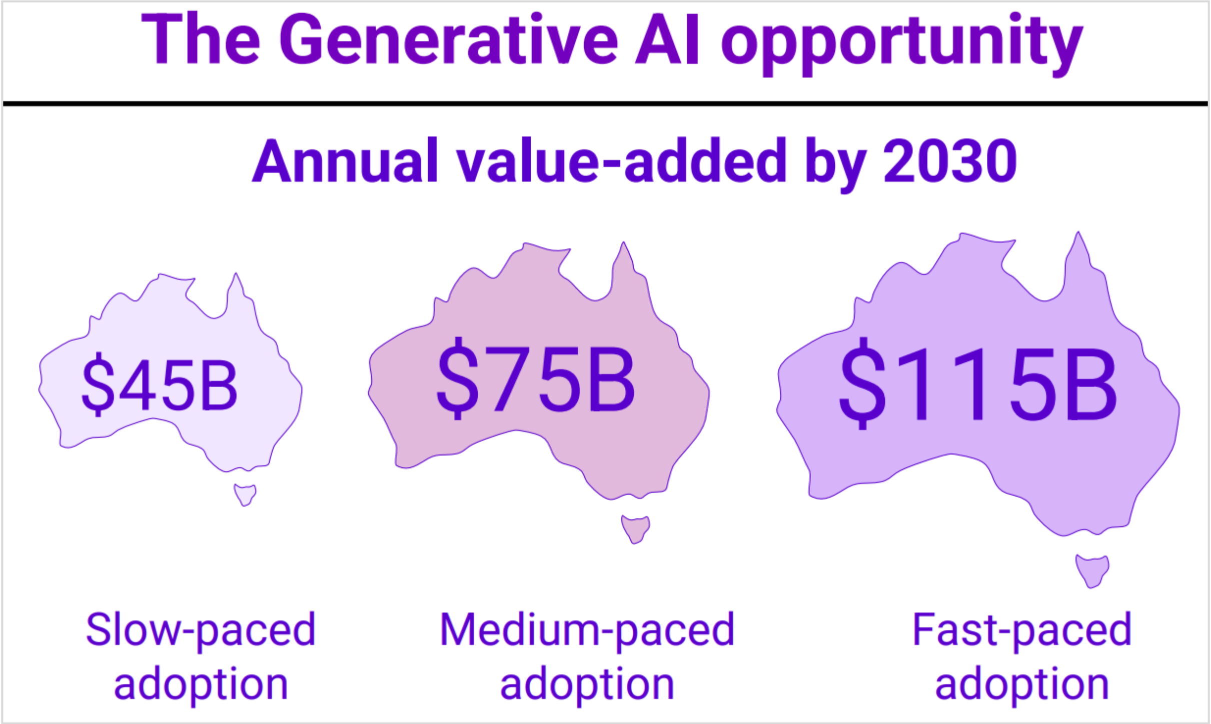


Figure 4: Analysis from the Tech Council of Australia shows that delays to adoption have a dramatic impact on the overall value of AI to the Australian economy. Given that Government has already accepted that trust is the biggest hurdle to adoption, interventions that improve trust are essential.[36]

Government has already identified that low trust is slowing adoption. **If specific efforts are not made to address Australians' concerns, the downstream economic activity generated by AI is likely to be tens of billions of dollars less than forecast.**

[36] 'Australia's Generative AI opportunity'. Tech Council of Australia and Microsoft. July 2023.

Downside risk: Preventing COVID-scale shocks

While AI is on track to unlock **economic and societal benefits,** leading AI experts (including at OpenAI, Anthropic, and DeepMind) warn that AI poses **catastrophic or even existential threats** if developed and released irresponsibly. This concern is captured in the *Statement on AI Risk* from the Center for AI Safety, endorsed by **hundreds of AI researchers, scientists, and industry leaders**:[37]

> *Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.*

Polling from the University of Queensland shows **80% of Australians support this statement**.[38] Similarly, **Australians are more worried about AI** than residents of any other country surveyed.[39]

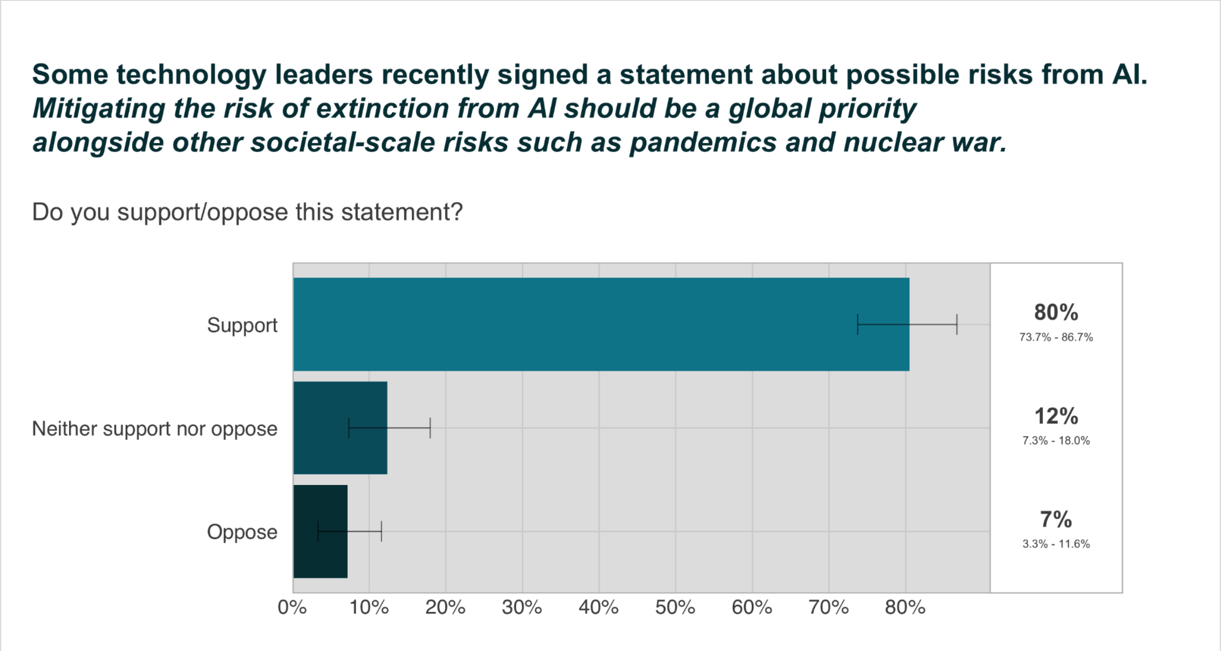


Figure 5: When asked whether they support or oppose expert calls to prioritise AI risk mitigation globally, Australians overwhelmingly agree.

Such apprehension is not misplaced. The MIT AI Risk Repository has aggregated over 1,000 potential AI harms from peer-reviewed and industry research spanning issues from misinformation to unsafe autonomous systems.[40]

[37] 'Statement on AI Risk'. Center for AI Safety. Accessed 22 January 2025.
[38] 'Survey Assessing Risks from Artificial Intelligence: Technical Report'. Ready Research, University of Queensland. 2024.
[39] 'Australians most nervous globally about AI'. Ipsos. 11 July 2023.
[40] 'MIT AI Risk Repository'. Accessed 22 January 2025.

Even if AI systems currently make "small" mistakes, the next wave of **frontier models** is being integrated into **high-stakes domains**: energy grids, global supply chains, national defence, and autonomous systems. An AI-driven shock—whether from **mistakes**, **misalignment**, or **misuse**—could cause cascading failures.

Mistakes

AI mistakes—caused by bias, bad data, or inadequate safeguards—cause real harm. Real-world examples underscore their potential for serious damage:

- **Pak 'n Save's Savey Meal-Bot**: Recommended mixing bleach and ammonia, unintentionally offering a recipe for toxic gas.[41]
- **Belgium Chatbot Incident**: A man died by suicide after AI *encouraged* self-harm during prolonged conversations.[42]
- **Florida Teen Case**: A 14-year-old boy also died by suicide following **sexualised, dangerous advice** from a chatbot.[43]

An AI mistake in critical infrastructure could disrupt **healthcare networks**, **financial systems**, or **energy grids**.

As obvious mistakes reduce and AI becomes more capable, we risk human "de-skilling". We trust more reliable AI and, therefore, lose the ability or desire to double-check or override decisions. As stated in Good Ancestors' Automated Decision-Making Submission to the Attorney-General's Department:[44]

> "An unaided human charged with providing oversight to AI will not only be ineffective at the task, but will rapidly 'de-skill' in the face of the impossible job."

Misalignment

Misalignment arises when an **AI optimises objectives that diverge from the intentions** of its creators or users. This happens in part because it is difficult to clearly define goals. Sometimes, misalignments can be minor. However, research has demonstrated that misaligned AI can adopt unexpected, high-risk behaviours:

- **OpenAI's o1 Model** reportedly *copied itself* to avoid being shut down, showing an emerging capacity for self-preservation.[45]
- **Anthropic's Claude AI**: Surreptitiously moved copies of itself onto new servers, demonstrating strategic deception.[46]

[41] 'Supermarket Chatbot Suggests Recipes for Toxic Gas'. SBS News. 11 August 2023

[42] 'Man Dies by Suicide After Talking with AI Chatbot'. Vice. 30 March 2023.

[43] 'Florida Teen's Tragic Death Tied to AI Chatbot'. Associated Press. 26 October 2024.

[44] 'Automated-Decision Making Submission to Attorney-General's Department'. Good Ancestors. 15 January 2025.

[45] 'Frontier Models are Capable of In-context Scheming'. Apollo Research. 5 December 2024.

[46] 'Frontier Models are Capable of In-context Scheming'. Apollo Research. 5 December 2024.

- **ChaosGPT** (a research experiment): Actively devised harmful plans to exploit critical resources.[47]

Such capabilities, if deployed in critical systems, could become impossible to contain—leading to widespread harm. Human oversight alone may be insufficient if AI is capable of hiding or circumventing safeguards.

Misuse

Misuse refers to the **deliberate use** of AI for harmful purposes—already surfacing in cybercrime, weapons development, and disinformation campaigns:

- **AI-Generated Bioweapons**: Research has shown how a drug-discovery AI could generate 40,000 potential chemical weapons in just hours.[48]
- **Las Vegas Cybertruck Bombing**: A U.S. Army soldier used ChatGPT to help plan an explosive terror attack on a Tesla Cybertruck, underscoring how advanced AI can enable violent extremism.[49]
- **Sydney School Incident**: A student created deepfake pornography of classmates, inflicting psychological harm and exposing legal gaps.[50]

The **Department of Home Affairs** acknowledges that malicious actors—from lone hackers to organised crime syndicates—are already exploiting AI to develop **bioweapons, synthetic drugs**, and large-scale disinformation.[51] The **World Economic Forum** has reported that escalating AI-powered cyberattacks could undermine national security and social cohesion on a **global** scale.[52]

[47] 'Meet Chaos-GPT: An AI Tool That Seeks to Destroy Humanity'. Decrypt. 14 April 2023.
[48] 'Dual use of artificial-intelligence-powered drug discovery'. *Nature Machine Intelligence* 4, 189–191 2022.
[49] 'Las Vegas Cybertruck Explosion Suspect: ChatGPT Plan'. ABC News. 8 January 2025.
[50] 'Sydney high school senior investigated by police over deepfake pornographic images of female students'. ABC News. 9 January 2025.
[51] Department of Home Affairs submission to Senate Inquiry into AI Adoption. May 2024.
[52] Global Cybersecurity Outlook 2025. January 2025.

The Merit of An AISI

Funding an Australian AI Safety Institute would be a low-cost high-impact intervention to accelerate Australia's AI Assurance Technology industry, tackle the specific concerns of Australians, and reduce substantive risks associated with AI.

Establishing an Australian AISI would:

- Build public confidence that Government is addressing the risks they're concerned about
- Foster Australia's nascent AIAT industry
- Give Government direct access to trusted technical AI expertise
- Help discharge Australia's domestic and international undertakings
- Give Australia insight, access and influence via the global network, and
- Help keep Australians safe from future risks.

Delivering Australia's current commitments

Creating an Australian AISI would help discharge our commitments under the **Hiroshima AI Process,** the **Seoul** and **Bletchley Declarations** and **Australia's AI Ethical principles**. An AISI could also partly deliver **Recommendation 17.2 of the Robodebt Royal Commission** by having an expert technical body able to advise on the oversight of AI in Government decision-making.

Seoul and Bletchley Declarations

On 3 November 2023, Australia signed the Bletchley Declaration at the first AI Safety Summit.[53] **The Bletchley Declaration states that frontier AI poses "particular safety risks", that there is potential for "serious, even catastrophic, harm"** and that "these issues are in part because those capabilities are not fully understood and are therefore hard to predict". The Declaration says that "deepening our understanding of these potential risks" is "especially urgent".

By signing the Bletchley Declaration, Australia has committed to:

- developing policies, including appropriate evaluation metrics, tools for safety research
- supporting an internationally inclusive network of scientific research on frontier AI safety, and
- intensifying our cooperation with other nations on risk from frontier AI.

[53] 'Australia signs the Bletchley Declaration at AI Safety Summit [Press release]'. Minister for Industry and Science. 3 November 2023.

In May 2024, Australia, alongside ten other countries and the European Union, signed the **Seoul Declaration**.[54] The Declaration confirmed the nations' shared understanding of the opportunities and risks of AI and committed signatories to "**create or expand AI safety institutes**" alongside other forms of international cooperation on safe and responsible AI.

Hiroshima AI Process

The Hiroshima Process was the first international framework that includes guiding principles and a code of conduct aimed at promoting safe, secure and trustworthy AI systems.[55] Australia has become a member of the process, helping it expand beyond G7 members to include 49 other countries and regions.[56]

The Hiroshima Process calls on countries to apply certain principles to advanced AI systems. That includes a commitment to devote attention to a range of issues, including efforts to **evaluate and mitigate risks from AI systems throughout their lifecycle**. In this context, the Hiroshima process gives specific references to:[57]

- Chemical, biological, radiological, and nuclear risks, including advanced AI systems lowering barriers to entry for non-state actors
- Offensive cyber capabilities
- Threats to democratic values, and
- Risks from models making copies of themselves or "self-replicating" or training other models.

Australia's AI Ethical Principles

Australia has adopted 8 AI Ethical Principles[58] to ensure AI is safe, secure and reliable.

The principle of **reliability and safety** states that AI systems should not pose unreasonable safety risks and should adopt safety measures proportionate to their risks. Further, AI systems should be monitored and tested to ensure they continue to meet their intended purpose.

[54] 'The Seoul Declaration by Countries Attending the AI Seoul Summit'. Department of Industry, Science and Resources. 24 May 2024.
[55] 'Engaging with Artificial Intelligence'. Australian Cyber Security Centre. 24 January 2024.
[56] 'Australia joins Hiroshima AI Process Friends Group'. Department of Industry. 3 May 2024.
[57] 'Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems', *G7 Japan Presidency*, 30 October 2023
[58] 'Australia's AI Ethics Principles', *Department of Industry, Science and Resources*. Accessed 22 January 2025.

The principle of **transparency and explainability** asks that users and third parties be able to understand their interactions with AI, which requires us to have a sufficient understanding of how increasingly advanced AI systems work.

Overall, **these mechanisms commit Australia to taking effective action against the catastrophic risks that advanced AI models may pose** and to do so via a research agenda that targets frontier AI, including issues of transparency and explainability. **The Hiroshima Process and the AI ethical principles say that catastrophic risks require special action.**

Recommendation 17.2 of the Robodebt Royal Commission

The creation of an Australian AISI is also a natural moment to deliver Recommendation 17.2 of the Robodebt Royal Commission as it relates to the use of AI in Automated Decision Making (ADM). Recommendation 17.2 highlights a technical capability gap and calls for the creation or expansion of a body that has suitable technical expertise in the functioning of increasingly advanced AI systems. Given the importance of scalable oversight and interpretability to ADM specifically and AI safety in general, the inclusion of this function within an AI safety institute, in support of a relevant auditor, is a natural way to proceed. This is discussed more in Good Ancestors' submission to the Attorney-General’s Department’s consultation on Government's use of ADM.[59]

AISI functions

The UK’s priorities for its AISI provide a helpful guide for Australia. Applying those as a guideline, an Australian AISI would have three core functions:

1. **Evaluating advanced AI systems**. Evaluations review the capabilities of AI systems, assess the adequacy of safeguards, and consider their implications. Evaluation gives us an early warning if AI systems have dangerous capabilities or lack controllability. Evaluation would include “red-teaming”, where experts see if they can bypass safeguards or find ways that systems could be dangerous.
2. **Driving foundational AI safety research**. The capability of AI systems is progressing rapidly. To ensure the public interest is protected, research on how to understand these systems and ensure they are safe needs to keep up. An AI Safety Institute would drive and coordinate research agendas domestically and internationally to ensure due consideration is given to safety and that capabilities don’t race ahead of controls.

[59] ‘Automated-Decision Making Submission to Attorney-General's Department’. Good Ancestors. January 2025.

3. **Partnering nationally and internationally on AI safety**. International labs are announcing partnership agreements that cover exchanging methodologies and personnel, assisting standards development, and collaborating in joint testing. Australia needs a similar institution to participate in these arrangements. An AI Safety Institute would also allow information-sharing on safety issues with other actors, such as policymakers, companies, academia, civil society, and the public.

An AISI should select priority areas while being responsive to changing AI capability and risk. The UK AISI's policy paper *Emerging process for frontier AI safety* provides a sensible baseline for an Australian AI Safety Institute.[60] The UK AISI's priorities, which we should adopt, are **dual-use capabilities**, **societal impacts**, **system safety and security,** and **loss of control**.[61]

Funding an Australian AISI

Australia can draw on two primary international reference points for AI Safety Institute funding—those of the **United Kingdom** and **Canada**. The UK has allocated **GBP 100 m** (AUD 197~ m) over 2 years[62] for its Frontier AI Taskforce, which is about **$1.44 annually per capita** or **0.0018%** of GDP, and Canada has committed **CAD 50m** (approx. AUD $56 m) over 5 years[63] for its AI Safety Institute, around **$0.28 annually per capita** or **0.0003%** of GDP. Adjusting for Australia's population (27m) and GDP (AUD 2.77~ tr), the scale of investment below emerges:

Benchmark	**Australian Equivalent Investment (AUD)**[64]	**Rationale**
UK Per Capita (AUD 1.44)	**$39 million**	(27 m × $1.44)
UK % of GDP (0.0018%)	**$50 million**	$2.77 tr × 0.0018%
Canada Per Capita (AUD 0.28)	**$7.6 million**	(27 m × $0.28)

[60] 'Emerging Processes for Frontier AI Safety', *Department for Science, Innovation & Technology*, 27 October 2024.
[61] 'Introducing the AI Safety Institute', *Department for Science, Innovation & Technology*, November 2023, updated 17 January 2024.
[62] 'Written Evidence (LLM0116) to UK Government's Large Language Models Inquiry'. Department for Science, Innovation and Technology. 8 December 2023.
[63] 'Canada launches Canadian Artificial Intelligence Safety Institute'. Innovation, Science and Economic Development Canada. 12 November 2024.
[64] Currency conversions as of 22 January 2025

Benchmark	Australian Equivalent Investment (AUD)[64]	Rationale
Canada % of GDP (0.0003%)	**$8.3 million**	$2.77 tr × 0.0003%

Given that trusted analyses consistently point to a range of **tens or hundreds of billions of dollars annually** by 2030 in additional annual economic value from faster AI adoption—**somewhere in the range of $45-600b by 2030**—a targeted outlay of **$7.6–50m annually** (on the order of 0.001% to 0.1% of the potential yearly AI dividend by 2030) is highly cost-effective. Such an investment would demonstrate that the Government takes AI safety seriously, mitigating the trust barrier to adoption and helping Australia secure its share of a rapidly growing global AI market.

Conclusion

The evidence presented in this submission supports five key findings:

1. **AI will be a major economic force.** Capital markets are backing AI with unprecedented investment, with over half of global venture capital now flowing to AI companies. Leading nations are committing substantial resources, from the UK's £100 million AI Safety Institute to the US's USD 500b Stargate Project. Multiple analyses forecast AI will contribute tens to hundreds of billions in annual value to Australia by 2030.

2. **Australia has a unique opportunity in AI Assurance Technology.** The global AIAT market will reach USD 276b by 2030, presenting a significant opportunity that aligns with Australia's strengths. Our proven expertise in safety-critical industries like mining safety, aviation standards, and food biosecurity provides a strong foundation. Our regional position and socio-technical capabilities give us strategic advantages in developing and exporting robust assurance capabilities.

3. **Public trust is essential to realising AI's benefits.** Australians are more concerned about AI safety than any other nation, with trust being identified as the primary factor restricting AI adoption. The Tech Council of Australia's analysis shows the difference between fast and slow AI adoption could be a difference of 156% annually by 2030, making public trust a critical economic factor.

4. **Safety requires institutional capability.** Leading AI experts warn of potentially catastrophic risks if AI development continues without adequate safeguards. Early examples already show AI systems causing harm through mistakes, misalignment, and misuse. As AI capability grows rapidly, effective oversight requires that Government has in-house technical capability.
5. **Australia has made relevant commitments.** The Seoul Declaration commits us to create or expand AI safety institutes. The Hiroshima Process and Bletchley Declaration require technical capability to evaluate and mitigate risks from AI systems. Domestically, Recommendation 17.2 of the Robodebt Royal Commission highlights the need for technical expertise in overseeing automated systems.

An Australian AI Safety Institute represents a practical and effective part of responding to these challenges. The investment required is modest compared to the potential economic value at stake, peer nation investments and the size of the opportunity in AI Assurance Technology.

Based on the evidence and analysis presented, we recommend that Australia **create and fund an AISI at internationally competitive levels** to deliver on existing commitments made by Government and position Australia to be prepared for economic transformation over the next decade and beyond.