Good Ancestors Policy

Australian-led regional AI Safety and Security Institute

Archana Atmakuri, Joshua Krook and Eddie Major

2024-25 AI Governance Summer Fellowship

May 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.

This project originates from the Good Ancestors AI Governance Summer Fellowship. The Fellowship, a collaborative initiative between Good Ancestors and Arcadia Impact's AI Governance Taskforce, integrates participants into focused research projects addressing critical challenges in AI safety policy and governance, including catastrophic risk. The paper reflects the research and opinion of the authors and not necessarily the policy recommendations of Good Ancestors or Arcadia Impact.

Our thanks go to the fellows who care so passionately about being good ancestors to future generations.

Good Ancestors Policy

Executive summary

Australia has publicly committed to advancing its artificial intelligence (AI) safety and security capability through international declarations such as the Bletchley and Seoul agreements. However, unlike its UK and US allies, Australia has not established a dedicated national *AI Safety and Security Institute.* This institutional gap is becoming increasingly urgent to address as AI capabilities evolve rapidly and the risk landscape becomes more unpredictable.

An *Australian AI Safety and Security Institute* (AISI) should be a central component of the *National AI Capability Plan,* which is currently in development. An AISI would provide the necessary architecture and resources to coordinate AI safety research collaboration across universities, CSIRO, government agencies, and private sectors while offering a whole-of-government focal point for identifying and responding to emerging risks. This includes a critical monitoring function for AI threats, similar to the Australian Cyber Security Centre's role in cybersecurity.

An AISI would require cross-government expertise and draw on capabilities from multiple domains: technical, diplomatic, legal, and national security. It would also play a significant international engagement role, ensuring Australia contributes to the global development of AI safety standards while remaining connected to cutting-edge technical evaluations of emerging systems.

Unlike other Seoul Declaration signatories, Australia is yet to establish an AISI. An additional value proposition demonstrating the broader value of Australia's involvement alongside other countries may spur action. This paper makes the case that a Regional AI Safety Initiative (RAISI) could provide that additional value.

As international leadership on AI safety remains fragmented, Australia has an opportunity to shape the agenda in the Asia-Pacific. Through a RAISI, modelled on Australia's SEA-PAC Cyber program, an AISI could help coordinate regional responses to AI risks, assist in capacity building, and support trusted evaluation pathways across Asia and the Pacific.

This paper recommends a three-phase approach to establishing AISI:

1. A grants program to coordinate and support existing AI technical capability.

2. The establishment of AISI as a national institute to lead evaluations, policy advice, and global collaboration.
3. Scaling the AISI regionally via RAISI, positioning Australia as a constructive leader in AI safety and governance.

Establishing an AISI would be a practical, nation-building step that strengthens Australia's AI capability, protects national interests, and positions us as a trusted voice in global AI governance. An ambitious government could launch these elements simultaneously.

Recommendations

To ensure an effective AI safety capability for the nation, we recommend the staged national rollout of an Australian AI Security Institute (AISI) aligning with the *National AI Capability Plan* and expanding toward Australia playing a leading role in AI safety throughout the Asia-Pacific region.

Phase 1: Coordinate and support Australia's AI safety capability through a national grants program

Deliver an initial AISI-branded grants program to coordinate and support AI safety research across the university and private sectors. This foundational phase would signal strategic intent and:

- rapidly build Australia's domestic AI safety capability by funding targeted technical and applied domain projects
- position the AISI as an influential 'maker and shaper' of Australia's AI technical safety capability.

Phase 2: Establish a dedicated Australian AI Security Institute (AISI)

Formally establish an AISI as a federally funded institute with dedicated technical staff, infrastructure, and strategic mandate. This phase would see the AISI:

- be the central government body for monitoring AI risks, advising on frontier AI, and coordinating responses across departments
- serve as a focal point for national security and critical infrastructure risks linked to AI
- lead the development of AI safety tools and benchmarks

- conduct independent evaluations of advanced AI systems, including pre-deployment model testing in collaboration with global partners
- act as Australia's representative in the International Network of AI Safety Institutes

Phase 3: Seek expanded international leadership through a Regional AI Safety Initiative (RAISI)

A RAISI, modelled on the *Southeast Asia and Pacific Cyber Program* (SEA-PAC Cyber), would project Australia's AI safety leadership into the Asia-Pacific while supporting regional cooperation and capacity building. This final phase would:

- establish Australia as the trusted regional leader for AI safety engagement, standards, and knowledge sharing.
- facilitate cross-border collaboration on AI technical and applied domain research, especially for countries with limited domestic capabilities.
- promote regional standards for AI evaluation and risk governance, leveraging AISI's work.
- enable Australia to host regional AI safety forums and contribute to inclusive global AI governance, bridging the Global North and the Asia-Pacific.

Good Ancestors Policy

Table of contents

Introduction

This paper describes the value proposition of an Australian-led regional institute dedicated to safety and security in artificial intelligence (AI). An AI Safety (and now Security)[1] Institute—commonly referred to as an AISI—is "a state backed organisation for advancing AI safety for public interest by examining, testing and evaluating AI systems."[2] The idea of establishing an AISI was first introduced at the inaugural AI Safety Summit, where a group of countries, including Australia, signed the Bletchley Declaration, affirming that "AI should be designed, developed, deployed, and used in a manner that is safe, human-centric, trustworthy, and responsible". This marked the beginning of the United Kingdom and the United States setting up their respective AI safety institutes, alongside the launch of the International Network for AI Safety.

Domestic AI safety institutes typically have three core objectives: advancing AI safety research (technical and non-technical); external engagement between various stakeholders including industry, academia, and government; and global engagement for capacity building through regional cooperation. As of 2025, the safety institutes are coordinating to develop robust methods for verifying and validating the safety and reliability of AI systems, including automated capability assessments, red-teaming, human uplift testing, AI agents evaluations, and certification processes.

The United Kingdom's AI Security Institute focuses on three core pillars: **test advanced AI systems** and inform policymakers about their risks; **foster collaboration across** stakeholders companies, governments, and the wider research community to mitigate risks and advance publicly beneficial research; and **strengthen AI development practices** and policy globally.[3]

The United States' AI Safety Institute focuses on three core areas with a similar vision to the UK: **advance the science** of AI safety; articulate, demonstrate, and disseminate the

[1] "Safety", in the branding of early AISIs and their specific work programs is largely synonymous with "security" in the sense of national security. Safety in this sense refers to large scale issues, like the ability of AI models to be misused to build bioweapons or harms of highly capable AI agents, and typically not "safety" in the sense of narrow safety concerns, like workplace health and safety. Preference for "safety" or "security" in branding is largely an issue of public communication. This paper does not take a position on the branding question.

[2] UK Government Department for Science, Innovation and Technology. (2023, November). *Introducing the AI Safety Institute*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute

[3] Ibid.

practices of AI safety; and **support various stakeholders including** institutions, communities, and coordination around AI safety.[4]

Australia actively engages with the international community and has been vocal about ensuring AI remains human-centric. Australia is a signatory to both the Bletchley Declaration and the Seoul Declaration, reinforcing its commitment to responsible and trustworthy AI. Public sentiment reflects this urgency—with around 69% of Australians expressing concerns about AI safety, the highest level of apprehension globally.[5]

Australians' concerns about AI, alongside the substantive risks, underscore the **need for an Australian AI Safety and Security Institute (AISI)**—not only to address domestic unease but also to position Australia as a regional leader. An AISI is a timely opportunity for Australia to leverage its strengths and take the lead in the Asia-Pacific, helping address shared regional risks, concerns, and harms associated with the rapid development of AI. An Australian AI Security Institute (AISI) would serve as a key part of responding to AI threats and mitigating some of the worst impacts to our national security, wellbeing, and way of life, while responding to AI opportunities and harnessing potential benefits.

Existing regional arrangements are insufficient and inconsistent with nations that have established AI safety and security institutes. An Australian AISI would dramatically increase Australia's capacity to respond to the changing technological landscape, benefiting national security, cybersecurity, and our economy. Unlike current arrangements, an AISI would be a 'one-stop shop' in the government capable of responding to threats, rogue and terrorist actors, creating a national forum for AI risk and technological talent.

From e-commerce and chatbots to the automation of creativity and professional services, we are facing a sustained period of disruption to jobs and economic activity. Almost every area of the economy will be affected by AI technologies, making this more akin to the Industrial Revolution than the dotcom boom or smartphone revolution.

---

[4] U.S. AI Safety Institute. (2024, May 21). *The United States Artificial Intelligence Safety Institute: Vision, mission, and strategic goals* (Version 1.0). National Institute of Standards and Technology. https://www.nist.gov/system/files/documents/2024/05/21/AISI-vision-21May2024.pdf

[5] Ipsos. (2023, July 11). *Australians most nervous globally about AI.* https://www.ipsos.com/en-au/australians-most-nervous-globally-about-ai

Complacency is not a sufficient response for the challenges we are facing in the months and years ahead.

Our region will require strong leadership, international partnerships, and a sufficient baseline of technical skill capacity to handle the challenges we now face. Already, we are seeing the cost of inaction. Among other shortcomings, Robodebt highlighted the widespread lack of technical skills and nous in government. This suggests Australia is robustly unprepared for the adoption of AI, which will bring more and more complex challenges. Meanwhile, national security and cybersecurity threats have continued to emerge with the rise of foreign AI systems from competing nations.

DeepSeek's 2025 surge into public awareness caught Australia off guard. The government's response was largely framed around foreign interference and did not capture how AI has different risks from other Chinese technologies previously handled with the foreign interference lens. This shows Australia's national security apparatus is not closely tracking AI developments and their implications. An AISI in Australia would immediately help identify and respond to changes in the landscape before they impact the economy. A rapid response capacity is essential as AI systems become increasingly sophisticated, threatening to destabilise national security, the job market and our way of life.

Culturally, Australia is at a crossroads: we can continue our hands-off approach, drafting non-binding ethical principles and largely ignoring AI safety risks, or we can start to 'get serious' about AI capacity in Australia and our region. Getting serious means coordinating relevant technological capacity via a dedicated AI safety initiative to secure Australia's interests in an increasingly dynamic and conflict-prone international environment.

Balancing AI's economic promise with risk management

The widespread adoption of AI would bring Australia significant economic benefits. Google forecasts Australia's AI opportunity at A**$290 billion in 2030**. Microsoft told the

Australian Senate Select Committee on Adopting AI that forecasts predict AI would create "200,000 new jobs and contribute up to $115 billion annually to our economy".[6]

The Kingston AI Group's economic report in 2024 forecast "$200 billion per year and the creation of an additional 150,000 jobs from 2023-2030."[7] According to Access Partnership, **US$3 trillion** of economic benefits are expected for businesses in key Asia-Pacific countries in 2030, if AI-powered products and solutions are adopted.[8]

Capital markets are backing AI with unprecedented investment, with over half of global venture capital now flowing to AI companies.[9] Leading nations and regions are committing substantial resources, from the UK's £100 million AI Security Institute, to the EU's €200 billion AI Continent Action Plan,[10] and the US's $500 billion private-sector Stargate Project.[11] Multiple analyses forecast that AI will contribute tens to hundreds of billions in annual value to Australia by 2030.[12]

The global AI assurance technology (AIAT) market will reach USD $276 billion by 2030,[13] 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.

---

[6] Parliament of Australia. (2024, August 16). *Senate Select Committee on Adopting Artificial Intelligence – Hansard*. https://www.aph.gov.au/Parliamentary_Business/Hansard/Hansard_Display?bid=committees/commsen/28289/&sid=0003

[7] Kingston AI Group. (2024, April 16). *Australia's AI imperative*. https://kingstonaigroup.org.au/news-and-publications/f/australias-ai-imperative

[8] Kaul, A., Wei, S. C., Ridhuan, N., & Goh, L. (2024, May 26). Economic impact report: Strengthening Singapore's AI leadership with Google. Access Partnership. https://accesspartnership.com/strengthening-singapores-ai-leadership-with-google/

[9] Irwin-Hunt, A. (2025, January 8). *AI dominates venture capital funding in 2024. fDi Intelligence*. https://www.fdiintelligence.com/content/41641e67-f00f-53c0-97cb-464b3a883062

[10] European Commission. (2025, April 9). *Shaping Europe's leadership in artificial intelligence with the AI Continent Action Plan*. https://commission.europa.eu/topics/eu-competitiveness/ai-continent_en

[11] OpenAI. (2025, January 21). *Announcing the Stargate Project*. https://openai.com/index/announcing-the-stargate-project

[12] Brookes, J. (2024, June 12). *Google forecasts Australia's AI opportunity at $290b in 2030. InnovationAus*. https://www.innovationaus.com/google-forecasts-australias-ai-opportunity-at-290b-in-2030/

[13] AI Assurance Technology Report Team. (2024, May). *Risk & reward: 2024 AI assurance technology market report (Version 1.0)*. https://www.aiat.report

Despite these economic promises, Australians are more concerned about AI safety than any other nation, with trust identified as the primary factor restricting AI adoption. **Failure to generate trust in AI in Australia could generate an economic cost of up to $70 billion a year.**[14] The Tech Council of Australia's analysis shows that the difference between fast and slow AI adoption could result in a 156% difference in annual economic value by 2030, making public trust a critical economic factor.[15]

Getting the AI boom wrong will cost the Australian economy. The lack of appropriate safety, safeguards, transparency and other features would result in a pushback by consumers and other countries in lawsuits, market behaviour and sentiment. This cost has already been made clear in lawsuits, fines and disciplinary proceedings in a range of industries, by a range of governments.

In Australia, Robodebt cost the Australian Government **$1.872 billion** in a court settlement with wrongfully indebted citizens.[16] As we learned from the Robodebt recommendations, centralised expertise is critical for risk management. Now is the time to do this for AI. As Robodebt made clear, current departmental arrangements have been insufficient in responding to complaints, errors and misguided decisions relating to automation.

In Europe, companies face fines or product bans if they produce unsafe AI. This directly affects Australian companies operating in Europe. Pursuing unsafe AI, irresponsible AI or AI lacking sufficient safety standards, privacy, transparency and accountability, will mean that our products are less commercially viable in markets with stronger regulatory regimes. For example, Europe's top court issued Google a **€2.4 billion** fine for abusing the market dominance of its shopping comparison feature, which is powered by AI algorithms.[17]

While benefits are important, there's also significant value in managing downside risks. AI experts warn of potentially catastrophic risks if AI development continues without

[14] Tech Council of Australia, & Microsoft. (2023). *Australia's generative AI opportunity.* https://techcouncil.com.au/wp-content/uploads/2023/07/230714-Australias-Gen-AI-Opportunity-Final-report-vF4.pdf

[15] Ibid.

[16] ABC News. (2022, August 26). *A Robodebt royal commission has been announced. Here's how we got to this point.* https://www.abc.net.au/news/2022-08-26/robodebt-royal-commission-explained/101374912

[17] Chan, K. (2024, September 10). *Google loses final EU court appeal against 2.4 billion euro fine in antitrust shopping case.* AP News. https://apnews.com/article/google-european-union-antitrust-shopping-court-a281e4e4722efa816e929a52a9939d86

adequate safeguards.[18] Early examples already show AI systems causing harm through mistakes, misalignment, and misuse. As AI capability grows rapidly, effective oversight requires that the government has in-house technical capability.

The Seoul Declaration commits us to create or expand AI safety institutes.[19] The Hiroshima AI Process[20] and Bletchley Declaration requires technical capability to evaluate and mitigate risks from AI systems.[21] Domestically, recommendation 17.2 of the Robodebt Royal Commission highlights the need for technical expertise in overseeing automated systems.[22]

[18] Roose, K. (2023, May 30). *A.I. poses 'risk of extinction,' industry leaders warn*. The New York Times. https://www.nytimes.com/2023/05/30/technology/ai-threat-warning.html

[19] Ministry of Foreign Affairs, Republic of Korea. (2024, May 23). *Seoul Declaration for safe, innovative and inclusive AI by participants attending the leaders' session of the AI Seoul Summit, 21st May 2024*. https://www.mofa.go.kr/eng/brd/m_5674/view.do?page=1&seq=321007

[20] Ministry of Foreign Affairs of Japan. (2023, October). *Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI Systems*. https://www.mofa.go.jp/files/100573471.pdf

[21] Department of Industry, Science and Resources. (2023, November 2). *The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 November 2023*. https://www.industry.gov.au/publications/bletchley-declaration-countries-attending-ai-safety-summit-1-2-november-2023

[22] Royal Commission into the Robodebt Scheme. (2023, July 7). *Report of the Royal Commission into the Robodebt Scheme*. https://robodebt.royalcommission.gov.au/publications/report

Acknowledging unprecedented security risks

Australia faces emerging threats from advanced AI systems with capabilities for misuse, deception, persuasion, and manipulation that current Australian frameworks are ill-equipped to address. Security risks are acute for systems developed by foreign governments or companies that are misaligned with, or disregard, Australian interests and values. Security risks also emerge from systems that claim to reflect our interests or values until technical problems around safety and alignment are resolved. AI poses significant risks to Australia, including critical infrastructure, data sovereignty, and citizen safety.

The assumption of Western technological dominance in AI development is increasingly questionable. Major American technology companies acknowledge there is no 'moat' protecting Western AI industry leadership. This concern was validated in 2025 with the emergence of DeepSeek's R1, a Chinese AI model achieving comparable performance to Western models at a fraction of the development cost—millions rather than billions. DeepSeek's R1 questions the impact of trade embargoes on specialised processors as a strategy to limit China's progress in the field.

Australia faces multiple critical AI security threats that require a coordinated response, including:

1. **Economic and social manipulation:** Rogue actors, states, or adversaries could deploy AI agents to manipulate Australian markets, businesses and citizens, potentially resulting in significant revenue and tax losses or harm to democracy and social cohesion.
2. **Terrorist and criminal exploitation:** Terrorist groups and criminal organisations are already using AIs for recruitment, with sophisticated chatbots designed to gain trust and promote membership.[23] Currently, no systematic plan exists to counter this threat.
3. **Advanced biological threats:** AI systems may soon facilitate the development of novel biological warfare agents—threats that, by definition, lack established

[23] Weimann, G., Pack, A. T., Sulciner, R., Scheinin, J., Rapaport, G., & Diaz, D. (2024, January). *Generating terror: The risks of generative AI exploitation*. Combating Terrorism Center at West Point. https://ctc.westpoint.edu/generating-terror-the-risks-of-generative-ai-exploitation

countermeasures.[24] These technologies could enable actors with no relevant skills to build biological weapons.

4. **Cybersecurity vulnerabilities:** Bad actors are already employing AI for data breaches, scams, hacking, and malware production. AI-powered scams demonstrate unprecedented sophistication through deepfakes, synthetic audio, and flawless text generation. Despite being one of the most targeted countries for such scams, Australia lacks a comprehensive AI scam strategy. According to IBM's annual *Cost of a Data Breach Repor*t, the average Australian data breach reached A$4.26 million in 2024, representing a 27% increase since 2020.[25]

These are not the only risks from advanced AI systems. The *MIT AI Risk Repository* captures more than 1,600 risks and builds them into a taxonomy across seven domains.[26]

As AI development becomes more affordable and accessible, the number of entities capable of creating potentially harmful systems will grow. This 'democratisation'" of risk gives high-impact capabilities to a wider range of actors, from disaffected individuals to terrorist and paramilitary organisations.

Government intervention becomes essential as self-regulation fails

Major tech companies in the US and UK have shifted their focus away from AI safety and security,[27] with multiple rounds of job cuts to these departments.[28] The age of industry self-governance is over. Tech companies will not self-regulate, even on issues of safety and harm to human users. Governments must fill this void.

---

[24] OpenAI. (2025, April 16). *OpenAI o3 and o4-mini system card.* https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf

[25] Sainsbury, M. (2024, August 6). *Data breaches are costing Australian organisations, IBM report reveals*. TechRepublic. https://www.techrepublic.com/article/ibm-data-breach-cost-report-australia/

[26] Slattery, P., Saeri, A. K., Grundy, E. A. C., Graham, J., Noetel, M., Uuk, R., Dao, J., Pour, S., Casper, S., & Thompson, N. (2024). *The AI Risk Repository: A comprehensive meta-review, database, and taxonomy of risks from artificial intelligence*. MIT FutureTech. https://airisk.mit.edu

[27] Smith, J. (2025, March 24). *Trump administration moves to deregulate AI industry*. *The New York Times*. https://www.nytimes.com/2025/03/24/technology/trump-ai-regulation.html

[28] Vaziri, A. (2025, February 27). *Google announces more layoffs as employees track cuts in crowdsourced document*. *San Francisco Chronicle*. https://www.sfchronicle.com/tech/article/google-layoffs-efficiency-measures-20192909.php

Google recently removed core planks of its 2018 responsible AI pledge, including a ban on AI for autonomous weapons and surveillance.[29] Social media service X (formerly Twitter) has laid off 80% of its engineers working on trust and safety, leading to the removal of essential moderation of violent and extremist content, disinformation and election interference.[30] OpenAI and other major AI companies have laid off various AI safety researchers, downsizing these departments in favour of more profitable areas. OpenAI's CEO, Sam Altman, was briefly terminated by its board, allegedly for failing to uphold safe AI standards.[31] New players in the AI race, including China's DeepSeek, similarly do not prioritise safety. The open-source or open-weights nature of some models may also make them more malleable to bad-faith third-party actors.

Some model developers are also adopting an extreme free speech stance in their products. Grok promotes itself as a "truth-seeking AI companion for unfiltered answers" with few content guardrails.[32] Earlier this year one user discovered it was able to produce detailed instructions for creating chemical weapons.[33]

Government laboratories and institutions have historically played vital roles in ensuring public safety across various domains. These include organisations like the Therapeutic Goods Administration (TGA) for medical products and the Australasian New Car Assessment Program (ANCAP) for vehicle safety standards. Internationally, bodies such as the International Civil Aviation Organization (ICAO) have established crucial safety frameworks that transcend national boundaries. These precedents offer valuable models for developing a systematic approach to AI security and safety.

---

[29] Dave, P., & Haskins, C. (2025, February 4). *Google lifts a ban on using its AI for weapons and surveillance. WIRED.* https://www.wired.com/story/google-responsible-ai-principles/

[30] Brewster, T. (2024, January 10). *Elon Musk fired 80% of Twitter/X engineers working on trust and safety. Forbes.* https://www.forbes.com/sites/thomasbrewster/2024/01/10/elon-musk-fired-80-per-cent-of-twitter-x-engineers-working-on-trust-and-safety

[31] Booth, H. (2024, September 17). *Why Sam Altman is leaving OpenAI's safety committee. TIME.* https://time.com/7022026/sam-altman-safety-committee

[32] Wiggers, K. (2025, February 17). *Elon Musk's xAI releases its latest flagship model, Grok 3. TechCrunch.* https://techcrunch.com/2025/02/17/elon-musks-ai-company-xai-releases-its-latest-flagship-ai-grok-3

[33] Al-Sibai, N. (2025, February 25). *Elon's Grok 3 AI provides "hundreds of pages of detailed instructions" on creating chemical weapons. Futurism.* https://futurism.com/elon-musk-grok-3-chemical-weapons

Case study: The DeepSeek AI shock

On 10 January 2025, Chinese startup *DeepSeek* released a new AI reasoning model (*R1*[34]) and accompanying AI chatbot service available via the web (*DeepSeek.com*) and phone apps. It quickly drew attention for its impressive performance compared to leading AI models from established tech companies.

DeepSeek claims it took just two months and less than US$6 million to build R1, a fraction of the cost of its competitor's models, although the figure is disputed.[35] That news caused an AI market shock, with nearly US$600 billion wiped from the stock price of AI chip manufacturer nVidia.[36]

Adding to the global interest, was DeepSeek's decision to release R1 as *open-weights*—making it available for use, modification, and redistribution under a permissive licence. Instead of using Deepseek's AI chatbot web service (*DeepSeek.com*) users can download versions of the R1 model, and run them on their own computer infrastructure. Scaled-down versions of R1 can even be installed on a standard laptop or smartphone.[37]

While DeepSeek models have safety features, they are not as robust as those found in AI models from US-based competitors.[38] DeepSeek is also heavily censored; it won't answer questions about sensitive topics like the Tiananmen Square Massacre, and aligns with the Chinese government's worldview, including its territorial claims.[39]

---

[34] DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., … Liang, W. (2025, January 22). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. https://doi.org/10.48550/arXiv.2501.12948

[35] Tran, J. (2025, January 31). *DeepSeek development cost probably 100 times the sticker price: fundie*. *The Australian Financial Review*. https://www.afr.com/markets/equity-markets/us5-6m-probably-100-times-that-says-fundie-of-deepseek-price-tag-20250131-p5l8mj

[36] Carew, S., Cooper, A., & Banerjee, A. (2025, January 28). *DeepSeek sparks AI stock selloff; Nvidia posts record market-cap loss*. *Reuters*. https://www.reuters.com/technology/chinas-deepseek-sets-off-ai-market-rout-2025-01-27

[37] Triggs, R. (2025, February 1). *How I installed DeepSeek on my phone with surprisingly good results*. *Android Authority*. https://www.androidauthority.com/install-deepseek-android-3521203/

[38] Bhattacharya, M. (2025, January 29). *Is DeepSeek's latest open-source R1 model secure?* *Netskope*. https://www.netskope.com/blog/is-deepseeks-latest-open-source-r1-model-secure

[39] Lu, D. (2025, January 28). *We tried out DeepSeek. It worked well, until we asked it about Tiananmen Square and Taiwan*. *The Guardian*. https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan

In addition to the various AI-specific safety concerns, broader cybersecurity risks with the DeepSeek.com web service presented challenges for Australia. Cybersecurity researchers claim DeepSeek's user login information is routed through computer infrastructure owned by a Chinese state-owned telecom company that is prohibited from operating in the United States.[40]

Several nations, including Australia, moved to ban DeepSeek from their government systems in early February. However, events leading up to the ban reveal a gap in Australia's AI safety capability. Media reported on 2 February 2025 that "understanding who makes decisions on the specific use of DeepSeek in the federal government has proved tricky"[41] and described how the Attorney-General's Department—which in April 2023 issued the directive to prohibit the use of Chinese social media app TikTok on government devices[42]—had "referred enquiries to the Digital Transformation Agency, which in turn referred enquiries to the Department of Home Affairs", which two days later issued the direction to identify and remove "DeepSeek products, applications and web services on all Australian Government systems and mobile devices".[43]

The DeepSeek case demonstrates that Australia currently lacks an organised capability to monitor and assess emergent AI risks and coordinate timely responses, revealing the need for a for-purpose standing national capability. There will likely be more DeepSeek-style AI safety shocks in the coming months.

[40] Tucker, P. (2025, January 27). *Researchers link DeepSeek's blockbuster chatbot to Chinese telecom banned from doing business in US*. *AP News*. https://apnews.com/article/deepseek-china-generative-ai-internet-security-concerns-c52562f8c4760a81c4f76bc5fbdebad0

[41] Taylor, J. (2025, February 1). *As DeepSeek upends the AI industry, one group is urging Australia to embrace the opportunity*. *The Guardian*. https://www.theguardian.com/technology/2025/feb/02/as-deepseek-upends-the-ai-industry-one-group-is-urging-australia-to-embrace-the-opportunity

[42] SBS News. (2023, April 4). *Australia bans TikTok on all official devices*. https://www.sbs.com.au/news/article/australia-bans-tiktok-on-all-official-devices/fpn6uvlw7

[43] Department of Home Affairs. (2025, February). *PSPF Direction 001-2025: DeepSeek products, applications and web services*. Australian Government. https://www.homeaffairs.gov.au

Building AI safety and security capabilities creates national advantage

Australia has a unique opportunity to address current and future AI risks by establishing an AISI that enhances national capabilities. While other countries have already created dedicated institutions for assessing and responding to advanced AI risks, Australia is still developing its national AI strategy. This timing advantage allows us to integrate AI safety into our capability agenda from the beginning, enabling a coordinated approach that fosters innovation and safeguards public interest.

Historically, Australia's national AI initiatives have been criticised by industry and academia as being piecemeal,[44] negligible,[45] and slow.[46] In response, the Australian Government announced in December 2024 that it would develop a *National AI Capability Plan* and expected to deliver it toward the end of 2025.[47] The plan provides a critical window to recognise AI safety as an element of Australia's broader technological and economic resilience.

The National AI Capability Plan will have four objectives: to grow investment, strengthen AI capabilities, boost AI skills, and secure Australia's economic resilience.[48] AI safety is a vital component to achieving them. AI safety and domestic AI capability are not synonymous, but the two are linked. Nations that have more developed domestic AI capabilities—through investment in AI research, innovation and adoption programs, and hardware infrastructure projects—have also taken leading roles in shaping AI safety globally.

In January 2025, the United Kingdom published its *AI Opportunities Action Plan*, a bold roadmap aimed at harnessing AI to boost innovation, create jobs and transform public services.[49] The plan calls on the UK Government to continue to support and grow its *AI*

[44] Kingston AI Group. (2023, August 3). *Domestic research critical for responsible AI in Australia*. https://kingstonaigroup.org.au/news-and-publications/f/domestic-research-critical-for-responsible-ai-in-australia
[45] Sadler, D. (2024, May 15). *"Profound disappointment" over budget's AI funding. Information Age.* https://ia.acs.org.au/article/2024/-profound-disappointment--over-budget-s-ai-funding.html
[46] Brookes, J. (2022, February 28). *Govt's 'AI Action Plan' is lacking action. InnovationAus*. https://www.innovationaus.com/govts-ai-action-plan-is-lacking-action
[47] Department of Industry, Science and Resources. (2024, December 16). *An Australian-first AI plan to boost capability*. https://www.minister.industry.gov.au/ministers/husic/media-releases/australian-first-ai-plan-boost-capability
[48] Ibid.
[49] UK Government Department for Science, Innovation and Technology. (2025, January 13). *AI Opportunities Action Plan*. https://www.gov.uk/government/publications/ai-opportunities-action-plan

Safety Institute—now renamed to *AI Security Institute*—(AISI) and credits the UK AISI's success in conducting pre-deployment AI model evaluations as a "significant and growing source of international influence for the UK" in providing clarity for how frontier AI models will be regulated.[50]

An AI Security Institute would amplify Australia's ecosystem

Australia's existing initiatives focused on AI research, adoption and responsible innovation are fragmented, underfunded and limited in scope. An AISI would fill a critical structural gap by providing national coordination on AI safety, linking technical expertise across universities, CSIRO, government, and private sectors. To be effective, it must be carefully designed to avoid duplication and focus on leveraging existing initiatives, and aligning efforts under a coherent national strategy.

Australia has an economy that is the 12th largest in the world, holds a 0.8% share of global gross domestic product (GDP), yet produces 1.6% of global publishing on AI topics and "publishes research in AI application domains at a faster rate than the global average".[51] That demonstrates Australia has a small, but active technical AI research ecosystem, which per-capita outperforms other nations and is well placed to increase its contribution to the global AI safety field.

AI research activity across Australia's university sector is large and growing, with five universities having AI topic outputs comprising between 10% and 18.8% of their total research activity in 2021.[52] In addition, Australia is also home to several specialist AI facilities, which include:

- *National AI Centre (NAIC)*: established in 2021 to support the growth of Australia's AI industry and AI adoption.
- *CSIRO's Data61*: the data and specialist digital arm of the national science agency, and home to one of the world's largest sites for AI and data science research expertise.[53]

[50] Ibid.

[51] Hajkowicz, S., Bratanova, A., Schleiger, E., & Naughtin, C. (2023). *Australia's artificial intelligence ecosystem: Catalysing an AI industry*. CSIRO. https://www.industry.gov.au/sites/default/files/2024-07/AI%20ecosystem%20report%20Mar%202023%20Catalysing%20an%20AI%20Industry%20PDF.pdf

[52] Ibid.

[53] Commonwealth Scientific and Industrial Research Organisation. (n.d.). *Data61 Business Unit*. CSIRO. https://www.csiro.au/en/about/people/business-units/data61

- *Responsible AI Research Centre*: established in 2024 to conduct fundamental research into safe and responsible AI.[54]

An AISI could have an immediate impact by coordinating and amplifying the skills and research programs already present across the academic and research sector. As demonstrated above, Australia already has considerable expertise, and it could be immediately boosted and focused through tools such as a grant program. An AI safety grant program could be launched rapidly and operate in parallel with the steps to formally establish an AISI.

Establishing an Australian AISI and its key domestic programs would call on cross-government expertise and capacity. These requirements would range from expert AI technical skills (CSIRO and the Department of Industry Science and Resources); international engagement, safeguards and non-proliferation in AI safety (Department of Foreign Affairs and Trade); and national security, cybersecurity and critical infrastructure security (Department of Home Affairs).

To respond rapidly to global AI risks, and operationalise the nation's AI safety capability, the Australian Government might consider framing its AI safety and security strategy as an initiative of the Department of Prime Minister and Cabinet (PM&C).[55] There are precedents for this: the Australian Government's Cyber Security Review of 2014 and its subsequent Cyber Security Strategy (2016) were led by PM&C, which held responsibility for cyber security matters until the establishment of the Department of Home Affairs in late 2017.[56]

Structurally, an AISI should be distinct from existing departments and agencies. This is for at least three reasons:

---

[54] Commonwealth Scientific and Industrial Research Organisation. (2024, December 9). *Landmark research centre positions Australia for a safe and responsible AI future*. https://www.csiro.au/en/news/all/news/2024/december/landmark-research-centre-positions-australia-for-a-safe-and-responsible-ai-future

[55] In 2023, the Kingston AI Group of professors proposed establishing an "Office for Artificial Intelligence that reports to the Department of the Prime Minister and Cabinet" which would "collect data and undertake cross-departmental planning, and be responsible for monitoring any AI specific regulations, and general regulations that pertain to AI." — Kingston AI Group. (2023, July 26). *Safe and responsible AI in Australia: Submission to the Department of Industry, Science and Resources.* https://consult.industry.gov.au/supporting-responsible-ai/submission/view/314

[56] Commonwealth of Australia. (2016). *Australia's cyber security strategy*. Department of the Prime Minister and Cabinet. https://www.homeaffairs.gov.au/cyber-security-subsite/Pages/2016-cyber-security-strategy.aspx

1. **Public legibility and credibility.** A key purpose of creating a safety institute is to build public confidence in AI. Giving the responsibilities of a safety institute to departments or agencies whose remit includes driving AI adoption or advancing AI capability would undermine the critical trust-building agenda. Established practice in other industries, including aviation safety, is that technical safety institutes should be appropriately distinct from other elements of government. For instance, in aviation safety, Australia's aviation safety regulator (CASA) is separate from the safety investigator (ATSB) so that the ATSB can build trusted relationships with industry and signal to the public that safety is a genuine priority. If the investigator is part of the regulator, industry may be reluctant to share information about risks for fear that the regulator may begin an enforcement action.
2. **Trusted engagement with leading labs.** An AISI will want to seek early access to frontier models to assist with safety evaluations before models are widely available. This information is commercially sensitive and leading labs will expect robust assurances about how it will be used and protected. AISI governance documents that prove its work does not include AI capability development, adoption, or commercialisation will assist it build trusted relationships. The head of the AISI should be able to say unequivocally to the public and the industry that their mission is safety and nothing else.
3. **Independent governance arrangements.** A range of specific governance options exist, including the UK's approach of its AISI reporting to its equivalent of the Department of Industry. A key feature of any model must be appropriate independent governance. An AISI cannot be bound by the same rules for using AI as the rest of the government. For instance, the government benefits from the signalling advantage of clearly banning products like DeepSeek. However, an AISI must be free to use such models for research and testing. If an AISI does not have appropriately independent governance arrangements, research into potentially risky AI models would be hampered by red tape while also muddying Government's best practice messaging.

A narrow-but-deep strategy delivers disproportionate value

The existing international AI safety institutes offer useful models for national AI safety programs, outlining core functions such as technical evaluation, standards development, and international coordination. However, Australia's smaller and more

fragmented AI ecosystem requires its own approach; one that focuses on strategic specialisation, coordinated leveraging of existing capabilities, and targeted contributions to global AI safety efforts.

A strategy of identifying and pursuing narrow areas of AI safety expertise should not be viewed as a limitation, but is a recognised strategic opportunity. In early 2023, a group of Australia's leading AI researchers called on the federal government to pursue small-data 'nimble AI' as the nation's strategy.[57] Likewise, in November 2023 South Australia's parliament recognised that it was "not realistic to attempt to compete broadly with the AI activity of the US and China" and that it was the state's competitive advantage that it had "instead selected narrow, focused areas of specialised AI in which to excel."[58]

The 'narrow-but-deep' competitive advantage applies not only to our national AI research and domestic capability, but equally to our national AI safety positioning. Therefore, it is reasonable that Australia should pursue a strategy of narrow technical AI safety expertise, and strategic collaboration to fill small but critical gaps in the AI safety capabilities of our global allies.

Australia must specialise to succeed

Australia should prioritise developing world-class expertise in specific AISI functions rather than attempting comprehensive coverage. This targeted investment would enable Australia to become an indispensable partner in global AI safety efforts, contributing unique capabilities that complement rather than duplicate the work of larger allies.

Australia already possesses valuable technical AI capability through CSIRO Data61's Software Engineering for AI team, which produces high-quality research and tools for AI technical safety. Recent outputs include taxonomies and frameworks for AI system evaluation, metrics catalogues for AI accountability, and reference architectures for designing AI agents.

[57] Kingston AI Group. (2023, February 22). *Statement by the Kingston AI Group*. https://kingstonaigroup.org.au/news-and-publications/f/statement-by-the-kingston-ai-group
[58] Parliament of South Australia. (2023, November 14). *Report of the Select Committee on Artificial Intelligence*. https://www.parliament.sa.gov.au/en/News/2023/11/14/Report-of-the-Select-Committee-on-Artificial-Intelligence

An Australian AISI could expand CSIRO Data61's baseline capability by coordinating deeper engagement with the university sector. While AI model safety evaluation publications may be less attractive to university researchers focused on career advancement through traditional academic channels, a structured grants program aligned with national and global AI safety priorities could appeal to early-career researchers with the required technical expertise. Through this approach, the AISI would function as a national AI safety research capability 'maker-and-shaper' while simultaneously supporting the development of Australia's next generation of AI talent.

Australia's already possesses rich domain expertise

Australia also has an opportunity to lead in applied domain AI safety research—understanding how AI safety principles and techniques apply in specific real-world applications. Australia already possesses the foundational requirements that enable it to be a leader in applied domain AI safety research: established research capacity; access to high-quality datasets; and mature regulatory environments with industries seeking safe AI adoption. Domains of AI safety research where Australia may seek to specialise include:

- **Healthcare:** Australia has extensive and centralised public health services, longitudinal public health datasets and associated administrative datasets, and a well developed medical and health care research capability.
- **Finance:** Australia's finance and banking system is well regulated, with deep capital markets (superannuation system), and a stable but concentrated banking system with the 'big four' banks being large by regional standards.
- **Education:** Australia's public university sector is known for its high quality education and research output, and is well regulated and largely standardised.

Applied safety efforts in these domains could lead to fruitful commercialisation opportunities, including for the AI assurance technology (AIAT) market discussed above.

Coordination at home enables leadership abroad

An AISI should serve as the lead agency for AI safety advice across the federal government, coordinating a whole-of-government approach to AI risk management.

It would bring together technical, legal, security, and policy expertise from key agencies, enabling consistent guidance, monitoring and rapid response to new threats or

incidents. Mirroring the structure of cyber response coordination bodies, the AISI should maintain standing liaison roles across departments to ensure AI safety is treated as a cross-cutting national priority rather than a niche technical issue.

An AISI should act as Australia's formal representative within the International Network of AI Safety Institutes, fostering deep ties with counterparts such as the UK's AI Security Institute and the US AI Safety Institute.

This international engagement would ensure Australia has a 'seat at the table' in shaping frontier AI safety standards, while enabling knowledge exchange, shared evaluation protocols, and reciprocal access to models and data. These partnerships would also support Australia's ability to monitor, test, and respond to emerging risks in real time; this is particularly important given the speed at which global AI developments now occur.

Australia's pathway to becoming the Asia-Pacific's AI security and governance leader

Australia's AISI can establish a Regional AI Safety Initiative (RAISI) to lead the Asia-Pacific in responsible AI governance and critical infrastructure protection. RAISI would create opportunities for regional forums, partnerships, and innovation while delivering economic benefits.

Asia-Pacific AI governance is fragmented

The Asia-Pacific region is rapidly advancing in AI development, with various nations taking steps toward responsible implementation and regulation. ASEAN has published an Expanded Guide on AI Governance and Ethics for Generative AI, though this remains voluntary and lacks comprehensive risk assessment frameworks. Meanwhile, Pacific Island Countries are at a nascent stage of AI adoption, highlighting significant regional disparities in AI safety discussions.

Despite these initiatives, few national governments have passed AI legislation or developed independent AI capabilities. Many regional efforts focus primarily on economic benefits through AI literacy and workforce development programs, with governance frameworks still in preliminary stages.

Underrepresented regions need AI safety coordination

Given current geopolitical instability, moving beyond a Western-centric approach to AI safety is essential. Regional cooperation is urgently needed to address local and regional challenges, risks, and harms posed by AI.

Global AI governance bodies have yet to effectively incorporate the priorities and lived realities of countries from the global majority.[59] There is a clear call for support of regional AI safety institutes that can consolidate resources, develop local expertise, and advocate for underrepresented regions such as ASEAN and Pacific Island Countries in global AI governance discussions.

Australia has key advantages to lead regional initiatives

Australia is uniquely positioned to bridge disparities between the Global North and South in AI development through a multifaceted approach combining regional leadership with global collaboration. As the AI landscape evolves, Australia's role in fostering inclusive, ethical, and innovative AI practices grows increasingly significant on the global stage.

Australia's competitive advantages include a:

1. high-quality university research sector that attracts international talent and drives innovation in responsible AI development
2. foundation of ethical AI practices with a focus on cybersecurity and emerging technologies
3. growing AI sector projected to contribute over $280 billion to the economy by 2030, if trust is fostered, and
4. strong position as a key player in the Asia-Pacific region.

[59] Chowdhury, R. (2024, September 19). *What the global AI governance conversation misses*. Foreign Policy. https://foreignpolicy.com/2024/09/19/ai-governance-safety-global-majority-internet-access-regulation/

Australia's cybersecurity strategy provides a proven template

Australia can adapt its proven cybersecurity approaches to launch regional AI safety initiatives. The 2023-2030 Australian Cyber Security Strategy aims to position Australia as a leader by 2030 through six cyber shields to defend against threats. Key elements that could inform an AI safety approach include:

1. Clear strategic vision: Australia's cybersecurity strategy establishes specific goals, including Shield 6 focused on resilient regional and global leadership
2. Targeted capacity building: Programs like the Southeast Asia and Pacific Cyber Program (SEA-PAC Cyber) enhance regional capabilities, incident preparedness, and policy development with substantial funding ($43.2 million) for long-term sustainability
3. Inclusive multi-stakeholder engagement: Participatory frameworks like "Cyber ASEAN" involve governments, private sectors, and civil society from the outset, ensuring diverse stakeholder inclusion
4. Alignment with broader policy goals: Integration of gender equality, disability inclusion, and sustainable development goals ensures initiatives remain equitable and inclusive, and
5. Strong international partnerships: Collaboration with partners like the U.S. in Palau demonstrates how international cooperation can enhance regional capabilities by leveraging shared expertise.

Australian expertise can foster inclusive AI dialogue

As a middle power, Australia can wield significant influence in regional and international forums without being perceived as a hegemonic force. This positions Australia to shape global AI governance by bridging the gap between countries leading AI safety discussions and the Global Majority, fostering inclusive dialogue.

Australia's strengths in education, research, regulatory frameworks, regional leadership, and diplomatic experience enable it to connect AI superpowers with smaller nations, encouraging inclusive dialogue on AI safety and governance.

Domestically, Australia's engagement in shaping AI safety policies benefits its national interests while strengthening regional influence. Australia ranks among the top 10 countries for human-centric standards in responsible AI according to the Global Index on Responsible AI 2024, positioned above Singapore.[60] This commitment is reflected in the government's unified, human rights-first national approach to AI development and implementation.

A Regional AI Safety Initiative will coordinate Asia-Pacific responses

Drawing from Australia's successful cybersecurity approach, we propose creating a Regional AI Safety Initiative (RAISI)—modeled after the effective SEA-PAC program—to serve as a foundation for regional cooperation in AI safety and governance.

A RAISI would:

1. Enhance capacity building and Global South representation by strengthening regional expertise in AI safety and governance and providing a dedicated forum for Australia, ASEAN, and Pacific Island Countries to contribute to global AI governance discussions, moving beyond the current bilateral approaches
2. Facilitate cross-border exchanges among universities to advance AI safety research, bring together technical experts to address region-specific AI risks, and create and co-enforce regional standards for AI safety, particularly regarding data security and critical infrastructure management
3. Create a regional point of contact, leveraging Australia's strong diplomatic ties, stable political institutions, and solid technical capacity to enhance trust in evaluations and address region-specific challenges, and
4. Leverage existing Australian AI Safety Institute work, allowing countries to adopt established evaluations and recommendations, reducing duplication and streamlining AI oversight.

[60] Adams, R., Adeleke, F., Florido, A., de Magalhães Santos, L. G., Grossman, N., Junck, L., & Stone, K. (2024). *Global Index on Responsible AI 2024* (p. 21). Global Index on Responsible AI. https://www.global-index.ai/

Building partnerships for AI safety coordination

A RAISI built on the collaborative approach would seek to leverage connections with the highest-performing nations in our region, including Japan, South Korea, Singapore, and potentially China in a participant capacity. Building on the existing memorandum with Singapore, this initiative would create a shared pool of talent and resources to facilitate AI upskilling, training, and capacity-building across the Asia-Pacific.

RAISI could also work alongside the most relevant domestic technical institutes in Pacific countries with limited resources, helping them understand and communicate about advanced AI. This collaboration would ensure that region-specific risks are highlighted, AI's impact on less digitised societies in the Global South are properly assessed, and diverse perspectives are incorporated into socio-technical evaluations.

The United Kingdom's AI Security Institute serves as a model in this regard. After launching the inaugural AI Safety Summit in 2023, it successfully guided international and regional discussions on AI safety, bringing together leading AI powers. Notably, this included participation from Chinese delegates and contributed to the establishment of subsequent summits in Korea and France.

Australia could position itself similarly by hosting an AI Safety Summit that invites neighboring countries and regional partners to collaborate with the Australian government. This summit would focus on establishing regional standards, safeguards, and safety specifications for AI development and deployment. Beyond the immediate security benefits, such an initiative could generate significant trade and economic opportunities between participating nations.

Conclusion

By building on the existing initiatives, Australia can immediately make a difference nationally and regionally on AI safety. We recommend the staged national rollout of an Australian AI Safety and Security Institute (AISI) aligning with the *National AI Capability Plan* and expanding over time toward Australia playing a leading role in AI safety throughout the Asia-Pacific region.

The discussion and evidence in this report supports five findings:

1. **Australia must address AI risks to achieve AI's economic promise.** Australia has demonstrated commitment to AI safety through international agreements like the Bletchley and Seoul Declarations, yet faces the highest level of public concern about AI safety globally. Failure to generate trust in AI could cost Australia up to $70 billion annually, making public trust a critical economic factor.
2. **Establishing an AI Safety and Security Institute (AISI) will address unprecedented security risks and build national advantage.** The DeepSeek case study illustrates Australia's current unpreparedness for AI developments, while Robodebt demonstrates the consequences of insufficient technical capacity. An AISI would provide a centralised authority to coordinate technological capacity, systematically identify and respond to threats, and secure Australia's interests in an increasingly dynamic environment.
3. **A narrow-but-deep strategy for the AISI would deliver disproportionate value.** By focusing initial efforts on areas of Australia's existing strengths—such as expertise in safety-critical industries, standards, and AI agents—the AISI can establish itself as a credible voice in AI safety. This targeted approach allows Australia to maximise impact while building toward broader capabilities and ensures efficient use of resources.
4. **Domestic coordination will enable international leadership.** An AISI would serve as a whole-of-government focal point for identifying and responding to emerging risks while coordinating AI safety research collaboration across universities, CSIRO, and government agencies. This domestic coordination creates the foundation for Australia to contribute meaningfully to global AI safety standards and governance frameworks.
5. **Expanding the AISI into a Regional AI Safety Initiative (RAISI) will position Australia as an AI leader in the Asia-Pacific.** With regional AI governance currently fragmented and many nations underrepresented in global AI safety discussions, Australia has a unique opportunity to lead. By leveraging its existing cybersecurity strategy as a template and utilising Australian expertise to foster inclusive AI dialogue, the RAISI can help coordinate regional responses to AI risks, assist in capacity building, and support trusted evaluation pathways across Asia and the Pacific.