Good
Ancestors
Policy

Australia AI Policy, 2025-2028

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.

Good Ancestors is proud to help coordinate Australians for AI Safety. This paper does not necessarily reflect the views of all signatories to Australians for AI Safety statements.

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.

Australian AI Policy, 2025–2028

Executive Summary **4**

AI development in 2025 **5**

Artificial Intelligence is an 'Everyone' Issue 9

2025 Playbook **10**

Recommendation 1: Launch an Australian AI Safety Institute 10

Recommendation 2: Introduce an Australian AI Act 10

Would an Australian AI Act be bad for business? 11

Recommendation 3: Host the next AI Safety Summit 13

Recommendation 4: Attract global Australian talent for AI Expert Panel 13

Recommendation 5: Update the AI Safety Standard 14

Recommendation 6: Courageous industrial policy 14

AI development in 2026 and beyond **16**

Digital Assistance or "AI Agents" 16

The Changing Job Market 17

New risks from AI agents 17

AI mega-corporations & hidden capabilities 18

Escalating Digital and Physical Risks 18

Towards Superintelligence? 19

2026-2028 Playbook **20**

Recommendation 1: Future-Proof Australia 20

Whole of government focus on transformative AI 20

Scope future-proof tax and welfare reforms for an AI-driven economy 20

National AI Cybersecurity Uplift 21

Resilient regulatory frameworks 21

Recommendation 2: Pursue a Global AI Treaty to share benefits and manage risks 22

Executive Summary

AI capabilities leapt forward during the last government term. During this term, leading AI labs in the US or China may develop AI systems with human-like cognitive ability. Regardless, AI capability will continue to accelerate and will have wide-ranging impacts on everyday Australians.

Australia has the power to shape whether AI's impacts on the Australian economy and everyday Australians are positive or negative. Government should "clear the decks" of known and pressing AI issues in 2025 so it is ready to address the unpredictability that increasingly capable AI will bring in 2026 and beyond.

In 2025 Government should:

1. **Launch an Australian AI Safety Institute:** Launch an AISI to build domestic expertise, assess risks, develop tools, and aid regional partners.
2. **Introduce an Australian AI Act:** Introduce legislation, using regulations for flexibility on specifics like "high-risk" definitions, focusing on transparency, clarifying responsibilities, setting guardrails, and aligning with international standards.
3. **Host the next AI Safety Summit:** Attend the upcoming Indian Summit and bid to host the next, returning the agenda to core safety issues.
4. **Attract global Australian talent for the AI Expert Panel:** Proactively engage leading Australians overseas, instead of limiting selection to local applicants.
5. **Update the AI Safety Standard:** Revise the standard to provide practical guidance on advanced AI risks for both developers and deployers.
6. **Implement Courageous Industrial Policy:** Invest in AI compute infrastructure, leveraging Australia's energy advantages and supply chain opportunities so Australia has a toe-hold in a future AI-driven economy.

Depending on emerging evidence, in 2026-2028, Government should:

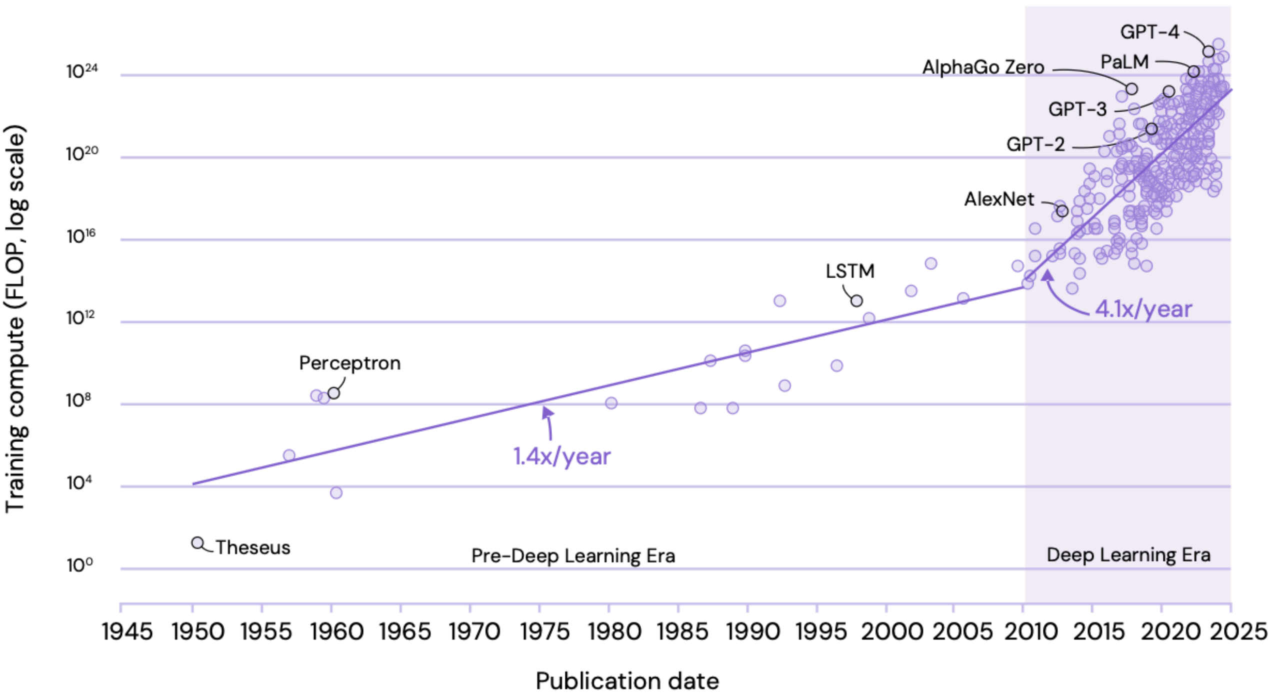
1. **Future proof Australia:** Get ready for powerful AI by:
   a. directing the public service to ready the country for AI risks and opportunities
   b. buttressing cybersecurity
   c. making legal frameworks resilient, and
   d. exploring readiness for an AI economy.
2. **Pursue Global AI Treaty:** Work internationally to establish AI capability ceilings, ensure global benefit sharing, and limit supranational corporate power.

AI development in 2025

AI progress will accelerate during this term of Government, bringing new opportunities, risks and disruptions.

President Trump's trade war won't slow AI. "Moore's Law", proposed in 1965, forecast chip growth through disruptions like the dot-com bubble. AI growth is subject to similar scaling laws. More and better-funded AI researchers working with more computing power and more data will continue to make increasingly powerful AI models despite hurdles.

More training compute was used for notable machine learning systems over time



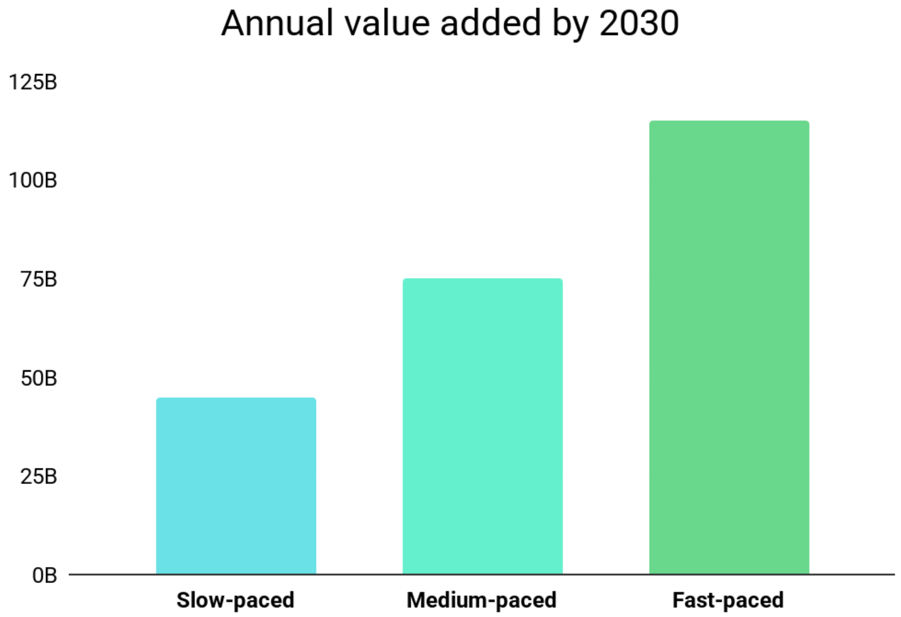
International AI Safety Report: As computing power increases, models perform better on practical tests. Performance also improves based on other factors, like the data used to train a model.

Markets forecast that Artificial General Intelligence (AGI) – AI models with human-like cognitive capabilities – will be developed during this term of government.[1] Google and OpenAI are calling on governments to prepare for a future with AGI. Anthropic forecasts AGI as early as 2026, but more likely in 2027. Even if these forecasts are optimistic, increasingly capable AI models will continue to shape society.

[1] Definitions of "AGI" are disputed. Often AGI means highly autonomous systems that outperform humans at most economically valuable work. Weaker definitions are limited to cognitive work while stronger definitions include embodied work. "Transformative AI" (TAI) often refers to AI systems with impacts similar to other general purpose technology like electricity or combustion engines.

> *Possibly by 2026 or 2027 (and almost certainly no later than 2030), the capabilities of AI systems will be best thought of as akin to… a 'country of geniuses in a datacenter'—with the profound economic, societal, and security implications.*
>
> — **Dario Amodei**, CEO, Anthropic

Even without AGI, low trust in AI is on track to cost Australia billions annually. Conservative analysis of generative AI's potential contribution to the Australian economy shows that fast AI adoption will be worth 2.6 times as much by 2030 compared to slow AI adoption. Policies that give Australians real reasons to have confidence in AI will close this gap.

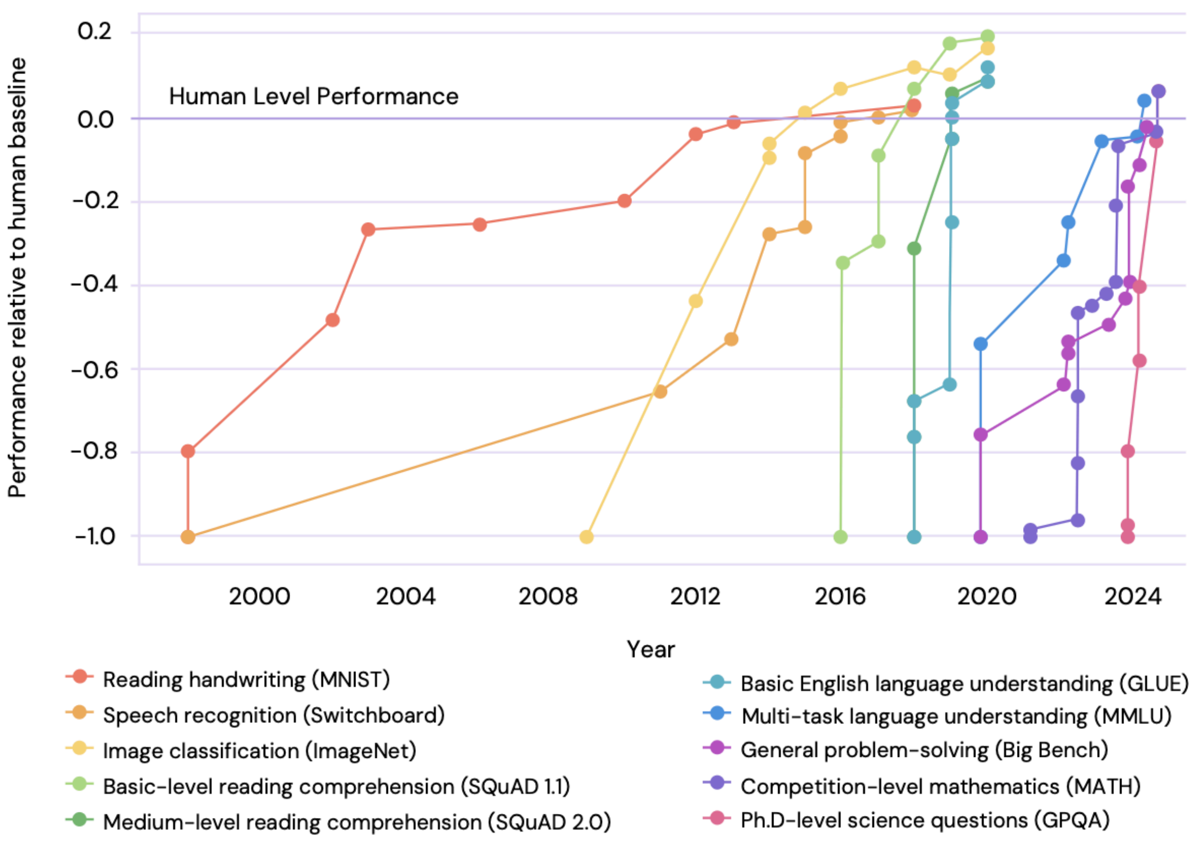


Analysis from the Tech Council of Australia shows that delays to adoption have a dramatic impact on the overall value of AI to the Australian economy (61% less value than fast-paced adoption). Given that Government has already accepted that trust is the biggest hurdle to adoption, interventions that improve trust are essential.

Australia must move quickly to establish a toehold in a world with highly capable AI. Australia needs to free itself to focus on the new issues that more capable AI creates, not be stuck clearing a "backlog" of known issues and ongoing projects. **Australia's priority for 2025 should be "clearing the decks" of known AI challenges so we can be ready for what will come in 2026 and beyond.**

AI is different from previous disruptive technologies. Because AI is software, it can spread quickly, simultaneously impacting many industries. Logistics—like building steam engines, cars or robots—slowed previous disruptions. With AI, companies can deploy thousands of copies almost overnight, leading to snap disruptions.

Further, AI is not like software we are familiar with. Traditional software operates on step-by-step instructions, whereas AI developed with machine learning is not written line-by-line, meaning understanding *why* it behaves a certain way or predicting its behaviours is challenging or impossible.



AI performance vs human performance on select benchmarks

International AI Safety Report: Scaling laws mean AI models rapidly approach and exceed human capabilities on a range of tasks. Experts tend to think that, ***if it can be measured, AI can solve it****.*

People outside the field are often surprised and alarmed to learn that ***we do not understand how our own AI creations work****. Powerful AI will shape humanity's destiny, and we deserve to understand our own creations before they radically transform our economy, our lives, and our future*

— **Dario Amodei**, CEO, Anthropic

From around 1910, cars quickly replaced horses as the primary mode of transport. Within fifteen years, horses went from doing most of the work moving people and goods to almost none. **Horses no longer made a valuable contribution to transport**. Humans re-skilled from supporting horses to supporting cars or into higher-skilled work. However, as AI surpasses humans in cognitive tasks, humans may become analogous to horses. Like horses, **there might be few cognitive skills where humans are more competent than AI alternatives**.

AI models doing a large portion of economically valuable work, say 10%, would lead to mass unemployment and the largest wealth shift in history. Roughly $10 trillion each year would move from countries like Australia to AI companies in the US or China. These companies would be more powerful than nations.

Artificial Intelligence is an 'Everyone' Issue

AI is rapidly evolving from a niche technology to a force impacting every aspect of society. As capabilities advance, its influence will become near-universal. While this paper concentrates on the **strategic and safety dimensions**, we recognise that **AI intersects with virtually every policy domain**, presenting challenges for governance and public discourse.

Challenges for governance

The broad impact of AI echoes historical challenges with cybersecurity. Governments grappled with whether cybersecurity capability should be centralised or distributed across all agencies. AI will be similar. The optimal approach likely involves embedding responsibility for understanding and managing AI within *all* relevant government agencies according to their remits. Distributed responsibility should be complemented by a central hub with expertise on frontier AI risks and opportunities, addressing those cross-cutting issues or novel challenges falling outside existing departmental mandates.

Challenges for public discourse

The breadth of AI's impact also complicates public discourse. Discussions focused on one specific AI issue (e.g., long-term safety) can sometimes be misinterpreted as dismissing the validity of other pressing concerns (e.g., near-term economic disruption or ethical implementation). Effectively navigating the AI transition requires acknowledging and addressing multiple valid concerns concurrently – the capacity to “walk and chew gum”.

While this paper prioritises strategic and safety considerations, we acknowledge the profound and interconnected implications AI holds for adoption, ethics, responsibility, employment, the environment, intellectual property, crime, and numerous other critical societal domains.

2025 Playbook

Australia is underway on a range of relevant work. In 2025, we should "clear the decks" of known challenges and current projects so we can focus on what will come in 2026 and beyond. We recommend 6 priorities for 2025:

Recommendation 1: Launch an Australian AI Safety Institute

Challenge: Australia committed to establishing an AI Safety Institute at the Seoul AI Summit and is the only signatory that has not fulfilled that commitment.

Currently, Australia relies on AI risk evaluations from foreign AI Safety Institutes or the internal labs of leading AI companies. While international collaboration is valuable, it cannot substitute for sovereign expertise. Australia needs sovereign technical capability to contribute to international networks, set its own strategic priorities, and verify external risk evaluations. Domestic capabilities to assess and mitigate AI risks are necessary to protect Australians and to help ensure that AI helps Australians thrive, rather than suffer new harms.

Recommendation: Australia is well-positioned to rapidly establish an AI Safety Institute (see **Attachment A**). Although there's no time to lose, it could be formally launched as a centrepiece of the Australian AI Safety Summit.

An Australian AISI would provide Australia with expert capability to assess AI risks, help our Pacific neighbours understand and manage risks, develop practical tools for overseeing AI systems, and build specialised expertise on emerging threats like AI agents. It would also strengthen public trust, boost the local AI assurance industry, and ensure Australia contributes meaningfully to global AI safety efforts.

Recommendation 2: Introduce an Australian AI Act

Challenge: Australian businesses face uncertainty about their responsibilities when deploying AI. The patchwork of existing regulators causes confusion and leaves gaps, like failing to cover risks from high-risk and general-purpose AI models or explain requirements for developers. Through its series of consultations, the Department of Industry, Science and Resources has established a solid foundation for an Australian AI Act, but no action has been taken.

Recommendation: An AI Act should be prepared for introduction as soon as practical.

To maintain flexibility, accompanying regulations can address key questions regarding the precise nature of "high-risk AI systems" and the specific regulatory obligations placed upon them. This approach allows for easier updates and alignment with international developments, such as those in the European Union. Focusing the AI Act on core issues like transparency and third party safety testing can help balance the need for meaningful safety measures while reducing risks to trade relationships, particularly with countries like the United States.

The AI Safety Standard discussed above can serve as a valuable reference point for determining what mandatory guardrails should look like in the AI Act.

An AI Act should respect the roles of existing regulators, including supporting them to understand and address the risks of AI systems and encouraging them to keep pace with the changing AI landscape. An AI Act should focus on new risks and risks that are not covered by existing regulators, like those from AI agents or specific dangerous capabilities.

Would an Australian AI Act be bad for business?

No. It would provide clarity and support AI use. Existing laws apply to AI, but how remains unclear. Model developers offer black-box systems and reject liability, leaving businesses uncertain of their obligations. This challenge intensifies as government and market pressures push them to adopt AI.

Other sectors have established laws and norms. For instance, car makers meet safety standards, businesses service fleet vehicles and provide driver training, and drivers obey road rules. When something goes wrong, it's typically straightforward to assign responsibility. For AI, there is little guidance or specialist support. Litigation has already started overseas, including a case where a chatbot caused serious harm. Australian businesses deploying chatbots may soon have to contend with similar litigation.

Impact of responsible AI policies in organizations, 2024

Source: McKinsey & Company Survey, 2024 | Chart: 2025 AI Index report

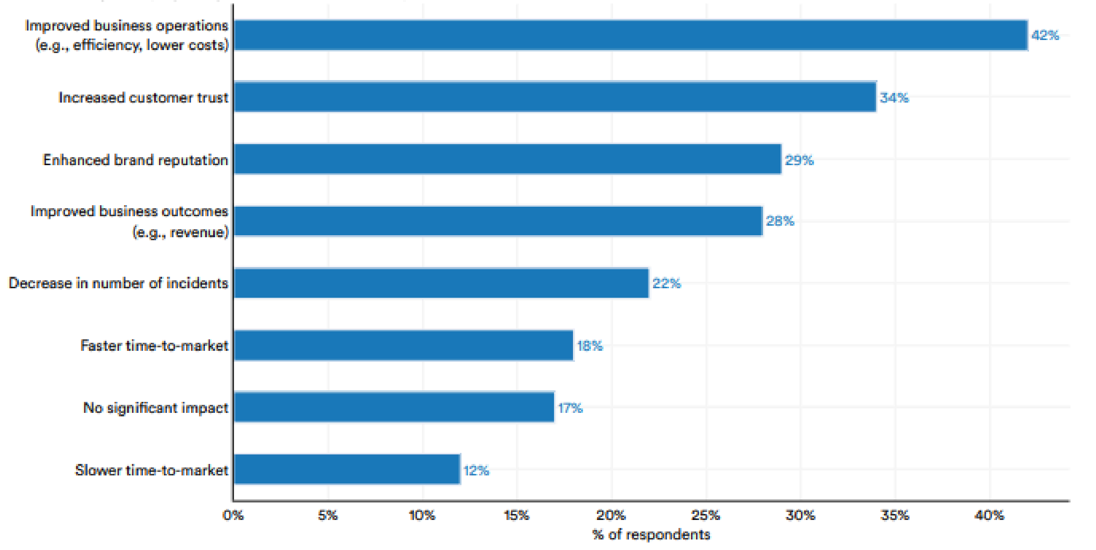


Figure 3.3.6[4]

Data show that responsible AI policy has few downsides for businesses and considerable upsides.

Benefits of an AI Act include:

- **Regulator clarity**: Require existing regulators (consumer, competition, privacy, etc.) to clarify how their rules apply to AI and issue timely guidance.
- **Defined responsibilities**: Provide rules outlining what developers, deployers, and end users must do, ideally preventing developers from offloading all liability, or at minimum providing certainty.
- **Guardrails and standards**: Introduce clear and practical technical and operational requirements for safe AI use, helping businesses understand and manage risks.
- **Liability frameworks**: Clarify when developers or deployers are accountable for harms, reducing legal uncertainty.
- **Transparency obligations**: Mandate developers publish risk evaluations and detailed reports on model performance and limitations. This helps Australian businesses understand what they're buying.
- **Harmonisation**: Align Australian AI rules with the EU to streamline global operations and enhance free trade.

Recommendation 3: Host the next AI Safety Summit

Challenge: Recent AI safety summits have focused more on opportunities than risks. While opportunities are exciting, market forces are already driving AI capability growth. The role of government, via safety summits, is to ensure that powerful technology is developed and used safely **despite** market forces.

Recommendation: Australia should attend the upcoming AI Safety Summit in India and bid to host the next summit.

Australia should propose an agenda focused on safety while creating appropriate places to discuss opportunities. Topics should include:

- The growing likelihood and shortening timelines for transformative AI.
- Specific safety and security risks linked to AI agents.
- Wider societal impacts, including the risk of wealth and power becoming concentrated in fewer hands.
- Practical steps to manage risks, including the urgent need for better AI oversight tools and effective global AI governance.

Recommendation 4: Attract global Australian talent for AI Expert Panel

Challenge: Brain-drain and self-selection mean Australia's current AI expert panel misses key perspectives. This is particularly the case for technical experts and those focused on transformative AI, who have largely moved to the US. Australian experts being out of step with global experts may partly explain why allies like the UK and Canada have moved faster to address AI risks than Australia. While Australia's current AI expert group includes top talent, like Bill Simpson-Young and Kimberlee Weatherall, important perspectives are missing.

Recommendation: Actively seeking the advice of Australians who have left the country would improve domestic AI policy. The new panel should *actively invite* Australian experts who have not worked with Government or are working overseas (like Toby Ord, Karl Berzins, Helen Toner, Huw Price, Peter Slattery, Cassidy Nelson, Michael Osborne or Daniel Murfet). The panel should also include leaders of non-profits and participants in the growing AI safety and security industry (like Harmony Intelligence, Cadent, Mileva Security Labs and Good Ancestors).

Recommendation 5: Update the AI Safety Standard

Challenge: Version one of Australia's Voluntary AI Safety Standard was criticised for not adequately addressing safety risks from advanced AI systems. While it provides a useful foundation, the Department of Industry has already conducted consultations on a second version.

Recommendation: An updated standard should move beyond general principles, be stronger on core safety points, and provide practical safety guidance. It should incorporate the recommendations in **Attachment B**, which cover:

- Providing specific guidance about how to use system cards and evaluation to select safer systems
- Enhancing testing and monitoring for dangerous capabilities, and
- Ensuring meaningful human oversight is supported by appropriate tools.

Recommendation 6: Courageous industrial policy

Challenge: Computing power could become the world's most valuable resource, but Australia has almost none of it. Computing power:

- Can allow AI systems to do economically valuable cognitive tasks anywhere in the world, and
- Help train next-generation AI systems.

Tariffs and supply chain disruptions are opportunities for Australia. U.S. chip diffusion rules restrict the global development of AI computing power. These export controls make Australia the best candidate in the Asia Pacific to centralise AI compute. Trade tensions also make Australia a more attractive destination for investment than other countries, potentially including the US.

The space is moving quickly, with OpenAI and the US working to partner with countries to build in-country data centres to "help support the sovereignty of a country's data" through the Startgate project. The US is also looking to reshape chip diffusion rules.

Recommendation: In the same way that Future Made in Australia is investing in hydrogen, quantum, and critical minerals, we should invest in AI compute to leverage our national advantages and secure our relevance in an AI world.

Advanced AI will likely have skyrocketing energy demands, which Australia is better placed to meet than most countries. Australia has the potential to be a global energy superpower, leveraging a net-zero economy and becoming a data centre leader. AI compute is a direct value-add to our natural energy resources. Australian data centres are already some of the most energy-efficient in the world. Mandala Partners estimates that Australian data centres in 2024 saved 2 Terawatt hours of electricity, a saving of 67% compared to on-premises compute.

When websites and apps with fast response times were key, Australia struggled to be a major data or compute hub because our geographical distance caused high latency. Even a half-second delay caused practical problems for users. However, this disadvantage is far less relevant to AI. Most uses of AI don't rely on low-latency connections to users. With generative AI, the time it takes for a response is not limited by internet speed, but by computation. Many of the most revolutionary uses, such as recursive research and development, can be done on-site, with results obtained days or weeks later. **For AI, a fraction of a second of latency is no longer relevant.** OpenAI already offers features like batch upload that provide a 50% discount for slower speeds, highlighting this trend.

While building new supercomputers would be expensive, it would become nearly impossible in the event of conflict over Taiwan. If that's the case, the supercomputers being built now might be the last supercomputers before Artificial General Intelligence (AGI) or Artificial Superintelligence (ASI) is invented. **Investment in compute today would be immediately economically valuable, would buy Australia a seat at the global table on AI, and could economically hedge against future global conflict.**

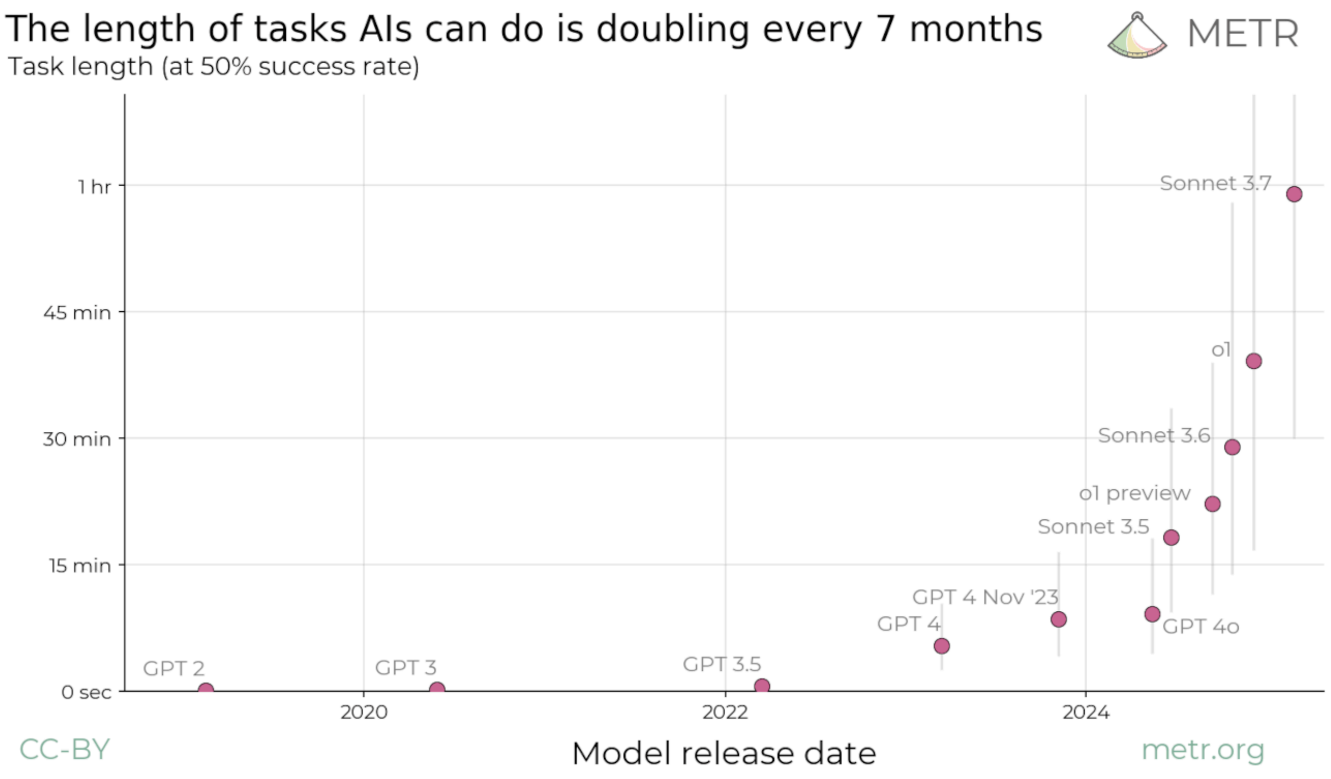
If we act now, Australia has a chance to become a global player. If we delay further, it will likely be impossible ever to catch up.

AI development in 2026 and beyond

Between 2026 and 2028, AI will transform our world faster than ever before. This is an exciting possibility and a significant challenge. The further into the future we forecast, the more uncertainty there is. But we can provide an evidence-based glimpse into what the future might hold:

Digital Assistance or "AI Agents"

Companies will release increasingly capable AI agents – programs designed to act independently online to achieve specific goals. By 2026, agents could manage projects, marketing campaigns, or analyse data as well as, or even better than, humans in some areas. While agents exist now, they are currently limited to simple tasks. Breakthrough agents will make headlines, but also spark widespread public concern about AI's rapid advancement. AI agents will have a day-to-day impact on Australians as they complete everyday tasks like making bookings, placing orders and completing paperwork.



AI agents were limited because they would lose the thread of longer and more complicated tasks. However, reliability has been increasing, allowing agents to complete *longer and more complex tasks. Beyond 2026, agents may be able to work for weeks or months without human intervention.*

The Changing Job Market

Increasing AI capability will shake up the job market. Recruiters might hire fewer people because AI can handle large parts of the workload for a growing number of positions. This could become a meaningful share of the economy by 2027. We could see software automating entire processes within industries like law or accounting, leading to sudden job losses as companies adopt AI systems that handle tasks accurately and efficiently.

When one company adopts these tools, competitors may feel forced to follow quickly, potentially leading to waves of restructuring across entire sectors. While new AI-related jobs will emerge, the speed of displacement in existing roles will likely outpace job creation, potentially pushing unemployment rates higher, perhaps even exceeding 10%. Unlike the spike in unemployment caused by COVID-19, there would be no obvious path back to “normal”.

Intellectual workers may struggle to reskill if AI surpasses human capabilities in many intellectual areas. When cars replaced horses, stablehands found new jobs, but horses didn’t. With powerful AI, humans are more like horses.

New risks from AI agents

As AI tools become more accessible, particularly powerful open-source versions, individuals might unleash AI agents with ambitious or disruptive goals – anything from "make me a million dollars" or "make this person famous" to spreading misinformation, generating deepfakes for scams, or orchestrating smear campaigns. Initial attempts might be clumsy, but could still create significant chaos. While this might sound far-fetched, there are rumours that an AI agent has already turned USD 50,000 of seed funding into USD 20 million, primarily through crypto trading and market manipulation.

As AI models improve, these uses could become more successful and harder to control, especially as regulators struggle to keep pace with freely available open-source technology being used for harmful purposes.

AI mega-corporations & hidden capabilities

We will hear increasingly frequent claims about new AI models being "Artificial General Intelligence" (AGI) – AI with human-like cognitive abilities – though experts will debate whether these claims are accurate. Regardless of the details, AI companies could rapidly become the most valuable corporations globally. Long-established businesses might face bankruptcy as AI erodes their competitive advantage. Law firms or investment banks will become less viable if AI models outperform most lawyers or bankers.

Leading AI firms might begin keeping their most powerful models private, using them "in-house" to perform profitable tasks rather than selling access and letting others profit. This could allow AI companies to identify industries ripe for automation, perhaps even strategically acquiring struggling companies, replacing the workforce with their internal AI, and rapidly expanding their dominance. This behaviour might outpace the ability of antitrust laws to respond. It also means the true cutting edge of AI capability could become hidden from public view, known only within powerful labs. If Australia has not established an AI Safety Institute, we will be blind to the AI frontier.

Escalating Digital and Physical Risks

Cybersecurity challenges will intensify. AI could write sophisticated malicious code, mimic voices in real time for scams (a significant step up from today’s "hi mum" scams), find security holes faster than defenders can fix them, and craft personalised, convincing fraudulent messages. This could lead to frequent, large-scale data breaches and significantly increase the cost of cybercrime to Australia's economy. Almost no Australians could protect themselves from a targeted cyber attack from a sophisticated actor.

Beyond the digital realm, reliance on AI in critical areas like healthcare systems, power grids, or supply chains carries risks. High-profile accidents or near-misses could occur if AI systems make unexpected errors under pressure. We’ll also see nations testing AI-driven military drones and robots, raising international concerns about new arms races. Terrorist groups and rogue states will explore using AI for dangerous purposes, such as designing bioweapons. OpenAI already says that its leading systems are “on the cusp” of being able to help novices make bioweapons.

Towards Superintelligence?

The most profound shifts will be when an AI lab announces that its current AI models are doing the majority of the research and development for their next generation of AI models. AI capability could progress at machine speed when AI matches humans in doing AI science. Rather than new models taking many months to develop, they could be developed in days or hours. This would lead experts to predict that Artificial *Super Intelligence* (ASI) – AI far exceeding human cognitive capability – might arrive in weeks or months, marking a pivotal moment for humanity.

A post-ASI world is hard to predict and could come this decade.

What about the "slowdown"?

Media reporting oscillates between highlighting the rapid development of AI and suggesting that AI capability has stalled.

While this reporting does not match the evidence, some systemic factors suggest a slowdown is possible. Most importantly, the AI scaling laws require increasing investment for steady capability growth. There's also a possibility that we run out of high-quality data.

A slowdown is possible, but unlikely to change Australia's AI policy priorities.

2026-2028 Playbook

What 2026-2028 looks like will be shaped by how quickly Australia moves in 2025. If Australia addresses pressing issues and adopts a courageous industrial policy we'll have options for exercising our sovereignty in the face of a changing world. If Australia remains slow to act, many options will be foreclosed, and we may have little say over our future as the world accelerates towards AGI.

Recommendation 1: Future-Proof Australia

Whole of government focus on transformative AI

Challenge: Powerful AI systems will have wide reaching impacts for every part of our society. While we can readily forecast some specific areas today, no think tank or expert can foresee the interaction between powerful AI and every dimension of society. Only a coordinated whole-of-government effort that takes the prospect of powerful AI seriously will be able to identify all the risks and opportunities.

Recommendation: Begin a coordinated effort across the Australian public service to build readiness for powerful AI. Although some global shocks are surprising and demand agility, we know that transformative AI is coming and we can start preparing now.

Scope future-proof tax and welfare reforms for an AI-driven economy

Challenge: Advanced AI could radically restructure the global economy, potentially leading to scenarios with very high structural unemployment and novel patterns of wealth distribution. Without proactive policy intervention, Australia risks seeing a large share of its GDP captured by overseas AI companies, while domestically, wealth concentrates among a small number of individuals, leaving a significant portion of the population unemployed and unemployable.

Government's rapid development of JobKeeper was impressive, but not without practical problems.

Recommendation: We should begin scoping tax and welfare reforms suitable for this potential future before transformative AI has significant real-world impacts. Options could include:

- Taxation mechanisms intended to ensure Australia retains a fair share of the value generated by AI within its economy, potentially limiting excessive wealth extraction by foreign entities.
- Taxes on AI-driven economic activity, such as on AI services or autonomous agents, potentially analogous to current payroll taxes but adapted for an automated economy.
- Welfare systems, possibly incorporating elements of Universal Basic Income, designed to provide economic security and dignity for potentially large numbers of citizens displaced from the traditional labour market.

National AI Cybersecurity Uplift

Challenge: Advanced AI is poised to fundamentally alter the cybersecurity landscape, likely favouring attackers who need only find a single vulnerability, while defenders must guard against all possibilities. AI agents could put the capabilities we currently ascribe to nation states into the hands of almost everyone.

In the longer term, the strategic posture may change if future developments, like AGI, enable the cost-effective creation of mathematically verifiable, secure code. This would shift the balance towards defence, allowing cybersecurity efforts to focus primarily on mitigating human factors.

Recommendation:

Australia faces a strategic choice:

- work internationally to prevent AI models with advanced offensive cyber capabilities from becoming widely accessible, or
- invest in sovereign Australian AI capabilities capable of actively defending all Australians in an AI-driven cyber environment (“AI vs AI”).

Prevention is typically better than treatment, but it would require immediate action. Preparation for an AI-driven cyber environment is likely prudent even if AI regulation reduces the risk of tools becoming widespread in the short term.

Resilient regulatory frameworks

Challenge: If AI systems automate significant portions (e.g., 10% or more) of valuable economic activity, the companies controlling these systems would generate revenue in trillions of dollars annually. Whether realised as profit or reinvested into developing even more powerful AI, this dynamic threatens to concentrate unprecedented economic and societal power within a few global AI corporations, potentially exceeding that of most or all nation-states.

Recommendation:

While passing an AI Act should be a goal for 2025, 2026 will have to involve making that AI Act and other regulatory systems robust to a world with powerful AI models.

Australian policymakers must consider how to design and implement robust AI regulations that remain effective under such conditions. This includes anticipating and building resilience against overwhelming lobbying, legal challenges, or informational influence campaigns funded by entities with vast resources, ensuring regulatory integrity and democratic processes are protected.

Recommendation 2: Pursue a Global AI Treaty to share benefits and manage risks

Challenge: The challenges of advanced AI transcend national borders. AGI will destabilise the global economy and geopolitics. Multi-trillion-dollar AI companies will become indispensable parts of the supply chain of every country and company, outstripping most or all countries in power and influence.

If AI companies direct Artificial General Intelligence (AGI) to work on the development of Artificial Super Intelligence (ASI), the future would become impossible to predict and could be catastrophic. While AGI has risks and opportunities, ASI offers few additional benefits and incalculable risks.

Recommendation: Australia should work with like-minded nations to negotiate a comprehensive global AI treaty. Building on concepts like those proposed for a Framework Convention on Global AI Challenges, this treaty should aim to:

- **Establish an AI Capability Ceiling:** Seek international agreement to prevent the development of AI systems that dramatically exceed human cognitive capabilities (approaching Artificial Super Intelligence), preserving human agency and control. A capability ceiling appeals to the rational self-interest of global leaders in maintaining sovereignty and stability.

    - This could be supported by a global AI safety agency, similar to the International Atomic Energy Agency or the International Civil Aviation Organisation.
- **Incorporate Global Benefit Sharing:** Design mechanisms to ensure the economic windfalls from AI are shared sufficiently to prevent extreme global inequality, mitigating the risk of destabilisation, particularly among nations possessing significant military capabilities (including nuclear weapons) who might otherwise face existential decline.
- **Limit Supranational AI Power:** Include provisions and enforcement mechanisms to prevent AI corporations from accumulating power that greatly eclipses the legitimate authority and regulatory capacity of nation-states.

Good
Ancestors
Policy

GRADIENT
INSTITUTE

Attachment A:

Proposal: Australian AI Safety and Security Institute

AI systems are rapidly becoming more capable, with promises of dramatic economic growth and progress on diverse challenges from health care to agriculture. However, AI opportunities are mirrored by AI risks. OpenAI warns that its models are "on the cusp" of enabling non-experts to build biological weapons, and autonomous AI agents may soon be able to conduct complex cyberattacks. To get the benefits, we must address the risks.

Governments have historically responded to analogous technological risks—like in aviation—by establishing independent technical bodies. For instance, the Australian Transport Safety Bureau's independence from regulators allows trusted, open sharing of safety-critical information between government, industry, and international partners.

Currently, Australia relies on AI risk evaluations from foreign AI Safety Institutes (AISIs) or the internal labs of leading AI companies (mostly in the US or China). While international collaboration is valuable, it cannot substitute for sovereign expertise. Australia needs sovereign technical capability to contribute to international networks, set its own strategic priorities, and verify external risk evaluations.

Global AI Risk Landscape

The global landscape for managing AI risks is deteriorating. Labs in the US and China operate without effective regulation and are reducing internal safety efforts. OpenAI recently ceased evaluating the risks of mass manipulation and disinformation. China's DeepSeek operates with minimal safeguards. Google and OpenAI are calling on governments to begin preparing for Artificial General Intelligence (AI with human-like cognitive ability). Anthropic's CEO predicts AGI in 2026/27. Markets predict late 2027.

The gap between global risk and domestic capability is growing. Domestic capabilities to assess and mitigate AI risks are necessary to protect Australians and to help ensure that AI helps Australians thrive, rather than suffer new harms.

An Australian AISI should perform four core functions:

1. **Sovereign evaluation of AI risks** Maintain the capability to verify claims from AI development labs, research organisations, and other AISIs. This includes assessing risks, distinguishing credible threats from exaggerated claims, and providing timely briefings to government agencies and organisations managing critical and societal infrastructure.
2. **AI oversight tools** Research and develop tools for effective human oversight and control of increasingly powerful AI models. "Humans in the loop" need tools that allow them to understand and direct AI. Safety engineers in other fields have technical tools that help them do their job. We need to explore the AI equivalents, including for use in government.
3. **AI agent research and risk management** Build specialised understanding of emerging threats from autonomous AI agents. Without new safeguards, these agents are predicted to be able to execute sophisticated cyberattacks, integrating technical capabilities (such as exploiting

‘zero-day’ vulnerabilities) with social manipulation tactics (such as spearphishing) as well as other kinds of threats. Existing AISIs are not yet focused on these risks.

4. **Regional engagement** Provide targeted support to Pacific neighbours who lack domestic capability in AI risk assessment. This includes offering rotating placements for regional experts, enabling them to develop expertise, raise national concerns, and advise their governments.

These functions leverage Australia’s domestic strengths and build an ongoing contribution to the international network of AISIs. Demonstrating meaningful contribution is crucial for ongoing access to international expertise and collaboration.

Comparable international AISIs spend between AUD 10–40 million annually, adjusted for population or GDP. Australia could deliver the proposed core functions for approximately **$15 million** per year.

This funding would cover:

- An agency head
- 10 expert technical staff
- 7 expert non-technical staff, including national security and threat assessment
- 3 Pacific Island placements
- 3 support staff
- Compute and technical resources
- Program funding for collaboration with industry and academia, and
- Operational costs.

New funding could be reduced to approximately **$10 million** per year by using seconded liaison officers, seconding DISR staff responsible for the AISI sharing network, and requesting DFAT to fund regional placements.

Securing World-Class Talent

Despite high demand for AI expertise, skilled professionals who understand advanced AI systems are often motivated by safety concerns. Many prefer working to address these risks rather than contributing to their development. Australians are already employed by overseas AISIs or independent AI safety organisations and may look to move.

With appropriate leadership and mission clarity, an Australian AISI would face minimal barriers in attracting top AI talent, including experts returning to Australia.

Rapid Launch—AI Safety Institute in a box?

Top Australian talent has clustered in charities, academic institutions, and start-ups:

- **Gradient Institute**, led by former senior staff from CSIRO’s Data61 and NICTA, and **Timaeus** employ leading Australian machine learning and math experts.
- **Mileva Security Labs** and **Good Ancestors** are led by individuals with experience in national security roles.

Australia could leverage existing pools of expertise to rapidly establish an AI Safety Institute. This approach could also attract leading AI talent back into Australia and back into government service. The UK took a similar approach to establishing its AISI.

Frequently Asked Questions

Does the National AI Centre (NAIC) or the Department fulfil this function?
The NAIC focuses on AI adoption and industry growth. While trust-building contributes to adoption, the NAIC's goals differ from managing AI safety risks. The Department currently engages with international AISIs and coordinates academia, but lacks the technical personnel and resources to evaluate emerging risks. A partner with a specific remit is also more likely to gain access to frontier models. Labs may be unwilling to share capabilities with agencies responsible for industry growth.

Should an AISI be part of a regulator?
Not initially. An AI regulator will require technical capabilities to assess AI products in the market—similar to how the Therapeutic Goods Administration operates labs for medical devices. However, an AISI addresses risks earlier in AI development, before products reach the market. Waiting to establish an AISI until legislation for a regulator is passed would delay critical action. Integration with a future regulator could be revisited during the regulator's establishment.

Is establishing an Australian AI Safety Institute a worthwhile economic investment?
Yes. Beyond managing risks from AI—such as enabling bioweapons or cyberattacks—an AISI creates economic opportunity. It would help catalyse Australia's AI Assurance Technology industry, positioning us to capture a share of a global market forecast to reach USD 276 billion by 2030. An AISI also addresses the 'trust gap'—Australians are more concerned about AI risks than any other nation, hindering adoption. Analysis by the Tech Council of Australia indicates that overcoming this trust barrier to enable faster AI adoption could unlock up to $70 billion per year in additional economic value for Australia by 2030.

What are the risks of creating an Australian AISI?
The primary risk is scope creep. AI presents diverse and evolving risks. The Bletchley AI Safety Summit highlighted frontier risks posed by advanced AI systems. Government must balance managing current and emerging threats. There are commercial incentives and existing regulators working to manage current risks from deployed AI systems. The risks of frontier AI systems represent a gap in the ecosystem that governments are best placed to address. An effective AISI must "ring fence" resources to focus explicitly on frontier AI risks. Attempting to address every AI risk or opportunity simultaneously would undermine its effectiveness.

Could Australian universities fulfil this role?
Partially. Universities have relevant expertise, including in sociotechnical risk and responsible AI. However, they lack trusted relationships with relevant stakeholders, including Departments, the national security community, and critical infrastructure operators. Universities are also typically not focused on practical understanding and response to emerging AI threats, and have limited ability to access frontier models. A dedicated institute would complement, not replace, academic expertise and include a budget for harnessing relevant talent in academia.

This document was prepared by Good Ancestors and Gradient Institute. Contact us for further details. 24/04/25

Attachment B:
Voluntary AI Safety Standard - Version 2

The Department of Industry has begun work to refine the Voluntary AI Safety Standard (VAISS). Although voluntary, VAISS v2 could inform future mandatory guardrails. This document proposes specific, practical improvements to strengthen the standard, focusing on better managing risks associated with advanced AI systems.

Developers and deployers have different roles with respect to making AI systems safe, and version 2 should separate those roles. The recommendations only cover some guardrails.

For AI Developers:

1. **Transparency:** Create and publicly share comprehensive "Safety Frameworks" outlining safety approaches and a detailed "Model Scorecard" for each model, detailing capabilities, limitations, and testing results.
2. **Evaluation Access:** Make models available for independent third-party evaluation and red-teaming.
3. **Model Security:** Implement robust cybersecurity measures to protect model weights from leaks or theft.
4. **Capability Control:** Test models for dangerous capabilities, withhold release if risks are unacceptable post-mitigation, and implement 'kill switches' where feasible to disable models if critical issues arise.
5. **Fair Risk Allocation:** Do not use Terms of Service (TOS) to transfer risks to deployers when they lack the practical ability to manage those risks.

For AI Deployers:

1. **Due Diligence:** Review developer Safety Frameworks and Model Scorecards before adoption. Prioritise developers demonstrating high transparency (e.g., via transparency indexes).
2. **Independent Verification:** Verify developer claims by seeking or reviewing independent, third-party evaluations of model performance and safety. Avoid models lacking adequate documentation or verification.
3. **Ongoing Monitoring Agility & Incident Response:** Document an AI incident response plan, including procedures for identifying and remediating AI incidents, communication protocols, and system deactivation if necessary. Be prepared to switch models or providers if risks are discovered.
4. **Human Oversight Tools:** Equip any 'human-in-the-loop' with effective and scalable tools necessary for meaningful oversight, beyond just assigning accountability.
5. **Contractual Awareness:** Carefully review developer TOS, particularly liability and reliability claims, ensuring alignment with the deployer's own legal obligations.
6. **AI Security Adaptation:** Recognise that AI security introduces risks distinct from traditional cybersecurity

Guardrail two

Risk Management: Create and use an ongoing process to identify, assess, and manage the potential risks and harms of AI systems based on their use.

Recommendations:

- Guardrail 2 currently highlights 'over-reliance on AI by users' as a key concept. Over-reliance is just one issue among many. The standard's risk management guidance would be stronger if it referred more broadly to the full range of known risks. Resources like the MIT AI Risk Repository offer comprehensive lists of these potential AI harms.
- Guardrail 2 requires organisations to have a risk management process. It should be improved by offering specific guidance on *how* organisations should assess AI safety risks, especially when using models developed by others.

  For AI deployers:

  - **Review the Developer's Safety Framework:** Deployers should check if the AI developer has a published 'Safety Framework' or similar document explaining their overall approach to safety (leading labs like Anthropic , OpenAI, and DeepMind provide these). This information should be a key input for a deployer's risk assessment. Deployers should exercise caution where a developer does not provide a safety framework.
  - **Review the Model Scorecard:** Deployers should review the specific 'Model Scorecard' (or equivalent) for the AI model they plan to use. This document should detail the model's capabilities, limitations, knowledge, behaviour traits, and safety testing results. Deployers should exercise caution where a model does not have a Model Scorecard.
  - **Verify Model Scorecard Adequacy and Accuracy through Third-Party Evaluations:** Deployers should seek independent, third-party evaluations rather than relying solely on developer claims. These evaluations are crucial for verifying the adequacy and accuracy of the Model Scorecard regarding performance and safety. Resources such as Stanford HELM, TrustLLM, and SEAL leaderboards can support this process.
  - **Make Informed Procurement Decisions:** Deployers should avoid using AI models that are not supported by a Safety Framework, a Model Scorecard, and credible third-party evaluations.

  For AI developers:

- AI Developers should produce a “safety framework” or equivalent document and make it available to the public.
- AI Developers should produce a “model scorecard” or equivalent document for each model they produce and make it available to the public in a timely manner.
- AI Developers should make their models available to third-party evaluators, including for red-teaming.

Guardrail three

System Protection & Data Governance: Protect AI systems through good cybersecurity and manage data properly, focusing on data quality, origin (provenance), and privacy.

During the Government consultations, participants highlighted that certain AI risks are already widely present. VAISS v2 should focus on risks where proactive measures can still be taken to prevent widespread issues.

One area where we can get ahead of the risk is the protection of model weights. If model weights are leaked, it is typically easy to remove safeguards. An unprotected version of the model can then become permanently available, limiting our ability to mitigate future misuse. This topic can be contentious because “open weight models” (like Qwen, Mistral, Llama or DeepSeek) offer both benefits and risks. One approach to reconciling this would be to ask developers to consider the risks their model poses prior to mitigations—if the mitigations are essential to moving the risk from acceptable to unacceptable, the model should be protected and not have open weights. Further detailed information and specific guidance on this topic are available from RAND.

- Developers should implement robust cybersecurity measures to protect their model weights effectively.

For AI deployers:

- Recognise that AI security introduces unique risks distinct from traditional cybersecurity (e.g., data poisoning, model evasion, prompt injection) and requires adapting security strategies to address vulnerabilities specific to AI models, data, and autonomous behaviour. Traditional security often focuses on infrastructure and software integrity, whereas AI security must also contend with risks inherent in the model's learning process, data inputs, and potential for unexpected behaviour.

Guardrail four

Testing & Monitoring: Thoroughly test AI models and systems against defined criteria before use, and continuously monitor their performance and behaviour once deployed.

Guardrail 4 (Testing) and Guardrail 2 (Risk Assessment) overlap. Considering risk assessment alongside testing could be a better structure for VAISS v2. For example, having developers submit their models for independent third-party evaluation is relevant to both testing and assessing overall risk.

For Developers:

- Developers should test their models to identify potentially dangerous capabilities. The results of these tests should be published transparently.
  - Models that still pose unacceptable risks after mitigation measures have been applied should not be released.
  - There are established procedures and known dangerous capabilities to evaluate. Resources like the Japan AISI: Guide to Evaluation Perspectives on AI Safety (Version 1.01) and the UK AISI: AI safety evaluations platform provide guidance. Ideally, an Australian AISI would contribute to this work.
- Where feasible, developers should maintain the ability to temporarily disable or shut down a model if a dangerous capability is discovered and safeguards cannot be immediately implemented. A 'kill switch' or other similar approaches can help address certain high-risk AI scenarios. The question of whether and how this applies to open-source products, particularly those with potentially dangerous capabilities, is an important consideration for mandatory or voluntary guardrails.

For Deployers:

- Deployers should review the 'Model Scorecard' (or equivalent) and any independent third-party evaluations for any AI model they are considering using.
- Where appropriate, deployers should engage an independent AI assurance company to conduct their own testing and evaluation. This would be particularly suitable if the AI system is considered moderate to high risk and where the organisation lacks sufficient internal technical expertise for evaluation. Utilising

Australia's growing AI assurance technology industry could also provide an incentive for domestic industry growth.

- Deployers should develop and document an AI incident response plan. The response plan should include procedures for identifying, reporting, investigating, containing, and remediating AI-related incidents such as severe bias discovery, security breaches, harmful content generation, or unexpected emergent behaviour. It should include clear communication protocols and mechanisms for human intervention. These should include:
    - The ability to comply with “recalls”, consistent with EU AI Act Article 79.
    - model or provider switching, (see below), and
    - System deactivation via “kill switch” (see above) if necessary.
- Deployers should be prepared to switch to a different model if a dangerous capability is discovered in the model they are currently using. This could involve changing to a different version from the same provider (e.g., moving from OpenAI's o1 to 4o) or switching to a model from a different provider (e.g., using Anthropic rather than OpenAI). A case study illustrating how a dangerous capability might be discovered in a widely used application, such as a chatbot used by an Australian small business (ranging from toxic behaviour to providing instructions for harmful activities), could effectively demonstrate the importance of this capability.

Guardrail five

Human Oversight: Ensure mechanisms are in place for humans to have meaningful control or intervene in AI system operations when needed.

Currently, 5.1.1 says that the “human in the loop” must have suitable competence as well as the necessary tools and resources. The next version of the Standard should expand on 5.1.1 by providing more detail about the necessary tools a person needs to oversee an AI system. The Standard should be explicit that a competent person alone cannot provide effective, practical oversight of a capable AI system. Just like in other fields, simply calling someone a safety inspector doesn't guarantee safety – they need to be equipped with the necessary tools to perform their inspection properly.

“**Scalable oversight**” tools and methods help humans oversee AI, especially for tasks too complex or rapid for effective or practical direct human supervision. The AI assurance technology industry is developing techniques that leverage AI interaction itself for oversight. One example is **AI safety via debate**, where AI models argue opposing viewpoints on a complex question, allowing a human to more easily identify the more truthful or sound positions. In general, tools create a scalable framework for guiding AI behaviour without requiring impractical human evaluation for every decision.

For Deployers:

- Deployers should ensure that humans involved in overseeing AI systems ('humans in the loop') are supported by appropriate and scalable oversight tools required to do their job effectively.

Guardrail eight

Supply Chain Transparency: Share necessary information about data, models, and systems with other organisations involved, so they can effectively manage risks.

There is some overlap between this guardrail and Guardrail 2 (Risk Assessment), as transparency is essential for conducting a thorough risk assessment. Some points related to transparency that were raised under Guardrail 2 could also be considered here.

For Developers:

- AI Developers should produce a "Safety Framework" or equivalent document and make it publicly available.
  - The safety framework should detail any procedural mechanisms for staff to raise risks, including details like how the organisation promotes a safety culture and whether non-retaliation procedures are in place.
- AI Developers should produce a "Model Scorecard" or equivalent document for each model they create and make it publicly available. High-risk and general-purpose AI systems may need more comprehensive scorecard.
- AI Developers should make their models available to independent third-party evaluators, including for 'red-teaming' (testing for vulnerabilities).
- AI Developers should not use their Terms of Service to shift risks to deployers or users where deployers or users have limited technical or practical ability to address those risks.

For Deployers:

- Deployers should consider safety frameworks and public "transparency indexes" when evaluating which AI developers to work with and generally prefer those who are more transparent and have better safety cultures. A good example of an index is the Foundation Model Transparency Index.
- Deployers should review the Terms of Service for any AI model they are considering using. Developers should pay particular to any claims made about reliability, quality, and accuracy, and to what extent liability is limited or excluded. Deployers need to be aware of their own legal obligations regarding reliability, quality, and accuracy, and understand that using a model without appropriate assurances might make it difficult to meet those obligations.