SCIENCE & TECHNOLOGY AUSTRALIA

POLICY SUBMISSION

4 OCTOBER 2024

PROPOSALS PAPER FOR INTRODUCING MANDATORY GUARDRAILS FOR AI IN HIGH-RISK SETTINGS

Science & Technology Australia thanks the Department of Industry, Science and Resources for the opportunity to respond to the Proposals Paper for introducing mandatory guardrails for AI in high-risk settings.

Science & Technology Australia is the peak body for the nation's science and technology sectors, representing 138 member organisations and more than 225,000 scientists and technologists. We connect science and technology with governments, business and the community to advance science's role in solving some of humanity's greatest challenges.

Key points

- Artificial Intelligence (AI) is having a transformative economy-wide effect globally that must be addressed in a centralised, coordinated and cohesive way to manage risks and realise its full potential.
- A human-centred approach is a prerequisite across all stages. This must be incorporated into AI legislation and regulation.
- An Australian AI Act that enshrines the principles for assessing high-risk AI and mandatory guardrails to address these, combined with a specific list of high-risk use cases would deliver the required certainty for research and industry, while also retaining flexibility to adapt to emerging AI developments.
- Australia's AI regulatory framework must clearly identify high-risk AI models and applications that and apply guardrails accordingly, while not stifling or limiting Australia's AI sector.
- Regulatory frameworks are just one component of ensuring Australia realises maximum benefit from AI – without a deep sovereign AI research capability, Australia risks falling behind globally and becoming reliant on other countries.
- Australia must also build a strong AI workforce with deep expertise and capability.
- Australia's AI development and deployment must be accompanied by measures to deploy clean energy technologies to supply the required computing power and data centres.

Science & Technology recommendations

1. The Australian Government commit to building Australia's sovereign AI capability, through establishing an Australian AI Agency that would oversee and support AI research, regulation, workforce development and industry engagement.

2. The Australian Government include environmental sustainability criteria in sovereign and high-risk AI development and governance to support efficient energy use, renewable energy policy and its net zero commitments.
3. The principles for assessing high-risk AI applications should be explicitly underpinned by a human-centric approach and include a guiding principle that all high-risk AI applications should retain human oversight on any decisions or outcomes from the application.
4. The principles for assessing high-risk AI applications should include a specific consideration of how an AI application may affect First Nations Australians, or draw on Indigenous knowledge or data.
5. The list of principles for assessing high-risk AI applications should be accompanied by a list of specific high-risk use cases to ensure certainty and consistency across AI development and deployment.
6. The full set of mandatory guardrails should only be applied to GPAI applications that are determined to be high risk. A requirement to ensure a human-centric approach, with appropriate human oversight, should be a requirement of all AI development and deployment.
7. To ensure a human-centric approach underpins all AI regulation, Guardrail 5 – *Enable human control or intervention in an AI system to achieve meaningful human oversight* – should be elevated to be the first guardrail.
8. An additional guardrail should be added that specifies that any AI development or deployment must involve human rights by design principles that will lead to deep and genuine engagement with First Nations people for any AI that includes Indigenous IP or affects First Nations Australians or communities.
9. To further strengthen accountability and transparency across the AI supply chain, the guardrails could be amended to require AI developers and deployers to develop and share longitudinal risk assessments.
10. The Australian Government should invest in Australia's AI research capability to enable deep understanding and inform high-quality assessment and regulation of GPAI to ensure its safety and efficacy from development to commercialisation stages.
11. To minimise the regulatory burden on small- to medium-sized businesses, the Australian Government should establish an Australian AI Agency to provide a single point of contact to support businesses navigate AI regulation requirements. The Government could also consider implementing a tiered approach to AI regulation and compliance, while still maintaining a mandatory focus on human-centric design and ensuring appropriate human oversight of any high-risk AI system.
12. The Australian Government should develop an Australian AI Act that enshrines the principles for assessing high-risk AI and sets out mandatory guardrails for high-risk applications, accompanied by a legislative instrument that lists specific high-risk use cases. The Act should also establish an Australian AI Agency to support and oversee Australia's AI sector, including regulation and compliance.
13. The Australian Government should consider programs and other ways to support the development of a deeply skilled AI workforce, including:
    - targeted measures to improve advanced maths participation levels
    - ways to incorporate a human rights by design approach to all AI development and deployment.

Australia must develop its sovereign AI capability

In considering high-risk AI, the Australian Government must fulfill its role in making sure AI is a net public good. While robust regulation is an essential component, it is not the only one. The biggest risk Australia is facing is failing to develop our own sovereign capability – we must ensure we do not become reliant on AI technologies developed in other countries that might not be suitable to Australia's unique communities and context. The Government must support the development of globally competitive sovereign AI capability, including a sustainable AI and digital infrastructure workforce.

Businesses have recognised the opportunities AI presents and are pursuing investing heavily. According to the Kingston AI Group, which consists of Australia's top AI scientists, greater use of AI in key Australian industries will lead to an additional $200 billion boost to GDP annually and 150,000 jobs from now until 2030. A 2024 Business Standards Institution study found that 67% of Australian businesses are expecting to increase AI investment this coming year, 60% believe it is important for employees to know how to deploy AI safely, ethically and effectively and 86% encourage employees to use AI. These factors highlight the increasing demand for AI – a rapidly evolving situation that is yet to be supported by appropriate and relevant governance and accountability frameworks.

Without sovereign capability and an Australian AI Act, Australia places itself at risk of anti-market behaviours such as market manipulation and exploitation, which, if left unchecked, may be unable to be curbed by the Australian Government. The five biggest companies in the world are AI companies. Global private AI investment expected to be at least $160b USD this year. With such heft, they may easily disregard any attempt to regulate and monitor activities which could harm everyday Australians. In the face of this potential dominance, an Australian AI Act can provide certainty about business obligations and will align industry with high standards of care without compromising their growth potential.

AI's environmental footprint should also be considered in developing regulatory frameworks and plans for Australia's AI capability. AI relies on data centres for storage and compute power. On a global level, data centres account for 1-1.5% of electricity use and are responsible for 1% of greenhouse gas emissions. The amount of compute power is significantly increase, with one estimate suggesting data centres could account for 14% of global emissions by 2040. Data centres also use vast amounts of water to keep equipment cool and running efficiently. Google's data centres increased their water consumption by 60% from 12.9 billion litres in 2019 to 21.1 billion litres in 2022. Australia's AI development must consider push for green energy should include environmental sustainability criteria into the assessment of sovereign and high-risk AI to minimise energy footprints.

One way to support Australia's sovereign AI capability and ensure a comprehensive and consistent approach to regulation is to establish an Australian AI Agency. This agency would oversee Australia's AI engagement, including building a robust sovereign AI capability, drive industry investment and research in AI, support international engagement and cooperation, build Australia's AI workforce and manage AI regulation and safety. An Australian AI Agency would also support strong policy capability to ensure Australia has a place in global discussions on regulatory regimes. This will be essential for Australia to keep pace with international developments and maintains a domestic framework that is relevant and interoperable.

Science & Technology Australia recommendation 1

The Australian Government commit to building Australia's sovereign AI capability, through establishing an Australian AI Agency that would oversee and support AI research, regulation, workforce development, industry engagement and global engagement.

Science & Technology Australia recommendation 2

The Australian Government include environmental sustainability criteria in sovereign and high-risk AI development and governance to support efficient energy use, renewable energy and net-zero policy commitments.

Consultation questions

Proposed principles for determining high-risk AI

In designating an AI system as high-risk due to its use, regard must be given to:
a. The risk of adverse impacts to an individual's rights recognised in Australian human rights law without justification, in addition to Australia's international human rights law obligations
b. The risk of adverse impacts to an individual's physical or mental health or safety
c. The risk of adverse legal effects, defamation or similarly significant effects on an individual
d. The risk of adverse impacts to groups of individuals or collective rights of cultural groups
e. The risk of adverse impacts to the broader Australian economy, society, environment and rule of law
f. The severity and extent of those adverse impacts outlined in principles (a) to (e) above.

1. **Do the proposed principles adequately capture high-risk AI? Are there any principles we should add or remove? Please identify any:**
   - **low-risk use cases that are unintentionally captured**
   - **categories of uses that should be treated separately, such as uses for defence or national security purposes.**

The proposed principles to help define high-risk AI and high-risk settings for AI set out a solid framework of considerations for assessing high-risk applications. However, it is not clear exactly how, on their own, they would determine whether an AI application is high-risk or not. A specific list of use cases – similar to those included in the Canadian and the European Union regulations – should be developed to sit alongside these principles to deliver clarity to the regulatory environment.

Additionally, missing from this list of principles is the imperative that all high-risk AI systems must retain a human in the loop, particularly high-risk applications. Much of the risk from potential AI applications and use cases will result from improper use, misapplication, bias or errors resulting from inadequate training data. As such, retaining humans in the loop to oversee and assess final decisions or outputs from AI applications is essential to avoid any potential adverse outcomes.

Defence and national security purposes will necessarily be treated differently, as they will have additional overlays of monitoring and security protocols. However, this does not mean they should be exempt from the principles outlined here, and most importantly, the imperative to keep a human in the loop.

Science & Technology Australia recommendation 3

The principles for assessing high-risk AI applications should be explicitly underpinned by a human-centric approach and include a guiding principle that all high-risk AI applications should retain human oversight on any decisions or outcomes from the application.

2. Do you have any suggestions for how the principles could better capture harms to First Nations people, communities and Country?

Building a robust Australian sovereign AI capability is essential to ensuring safe interactions with AI for First Nations Australians and communities. Reliance on AI applications developed overseas, rooted in other countries' cultures, populations, and environmental context, runs the very real risk of failing to account for Australia's unique situation – and particularly First Nations Australians' culture and knowledge. For example, using generative AI to source or write about the First Nations Australians runs the risk of the model sourcing information that is false, offensive and/or disregarding cultural protocols. This alone could lead to significant damage across all the proposed principles for First Nations people and other vulnerable communities.

To protect against this, and in line with an overall human-centric approach, the principles could include an explicit consideration that any AI developed that may affect First Nations Australians must be developed – or co-developed – with rigorous and in-depth consultation with First Nations communities.

Science & Technology Australia recommendation 4

The principles for assessing high-risk AI applications should include a specific consideration of how an AI application may affect First Nations Australians, or draw on Indigenous knowledge or data.

3. Do the proposed principles, supported by examples, give enough clarity and certainty on high-risk AI settings and high-risk AI models? Is a more defined approach, with a list of illustrative uses, needed?

- **If you prefer a list-based approach (similar to the EU and Canada), what use cases should we include? How can this list capture emerging uses of AI?**
- **If you prefer a principles-based approach, what should we address in guidance to give the greatest clarity?**

The principles are a good mechanism for assessing whether an AI application is potentially high risk, but they lack clarity for industry and other entities to operate with certainty and continuity. To deliver clarity across the various domains in which AI can be deployed, the principles should be accompanied by a list of specific use cases deemed to be high-risk (based on assessment against the principles), similar to the approach taken by Canada and the EU.

As discussed in subsequent sections, ideally, the principles would be set out in an Australian AI Act, with the specific use cases listed in a legislative instrument that could be updated as needed to address new developments or emerging applications.

Science & Technology Australia recommendation 5

The list of principles for assessing high-risk AI applications should be accompanied by a list of specific high-risk use cases to ensure certainty and consistency across AI development and deployment.

4. **Are there high-risk use cases that government should consider banning in its regulatory response (for example, where there is an unacceptable level of risk)? If so, how should we define these?**

AI developed in military contexts must be carefully controlled within a strong ethical framework. Australia should not develop AI to drive fully autonomous weapons systems. This is in line with the UN and International Committee of the Red Cross' call for countries to prohibit autonomous weapons systems unless they have human control in life and death situations.

5. **Are the proposed principles flexible enough to capture new and emerging forms of high-risk AI, such as general-purpose AI (GPAI)?**

Implementing a two-pronged approach of enshrining the principles (with the addition of ensuring the retention of humans in the loop) in primary legislation, coupled with listing specific use cases in a legislative instrument that can be updated as needed would deliver the required flexibility and agility to account for emerging applications and new AI developments.

6. **Should mandatory guardrails apply to all GPAI models?**

Applying all the mandatory guardrails to all GPAI models– regardless of risk level –would potentially be overreach. The process for applying mandatory guardrails should depend on whether the application is high-risk, as determined by applying the principles for assessing high-risk applications, and as specified in a potential list of high-risk use cases. However, the primary guardrail of ensuring a human-centric approach, with appropriate human oversight, should apply to all GPAI and other AI applications.

Science & Technology Australia recommendation 6

The full set of mandatory guardrails should only be applied to GPAI applications that are determined to be high risk. A requirement to ensure a human-centric approach, with appropriate human oversight, should be a requirement of all AI development and deployment.

7. **What are suitable indicators for defining GPAI models as high-risk? For example, is it enough to define GPAI as high-risk against the principles, or should it be based on technical capability such as FLOPS (e.g. 10^25 or 10^26 threshold), advice from a scientific panel, government or other indicators?**

Given compute capability is a proxy for the complexity and scope of a GPAI, it would also be reasonable to set a compute power threshold as a measure of how high-risk a model might be.

Proposed mandatory guardrails for high-risk AI

Organisations developing or deploying high-risk AI systems are required to:

1. Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance

2. Establish and implement a risk management process to identify and mitigate risks

3. Protect AI systems, and implement data governance measures to manage data quality and provenance

4. Test AI models and systems to evaluate model performance and monitor the system once deployed

5. Enable human control or intervention in an AI system to achieve meaningful human oversight

6. Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content

7. Establish processes for people impacted by AI systems to challenge use or outcomes
8. Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks
9. Keep and maintain records to allow third parties to assess compliance with guardrails
10. Undertake conformity assessments to demonstrate and certify compliance with the guardrails

8. Do the proposed mandatory guardrails appropriately mitigate the risks of AI used in high-risk settings? Are there any guardrails that we should add or remove?

The paramount imperative for AI to be developed and deployed safely is to ensure a human-centric approach and retain human oversight, particularly of any final decisions or outcomes. As such, Guardrail 5 should be elevated to the first and primary position in this set of protocols.

At a high level, these guardrails will support effective regulation of AI development and deployment. There will be further work needed at an implementation level to further strengthen these protocols to ensure they deliver a system that enables/supports effective regulation and public trust in AI and its use.

However, a clear-eyed approach must be taken to balancing these guardrails with the flexibility the industry and research sectors require to develop new leading-edge AI applications and ensure Australia is an attractive and competitive place to do both research and business.

Science & Technology Australia recommendation 7

To ensure a human-centric approach underpins all AI regulation, Guardrail 5 – *Enable human control or intervention in an AI system to achieve meaningful human oversight* – should be elevated to be the first guardrail.

9. How can the guardrails incorporate First Nations knowledge and cultural protocols to ensure AI systems are culturally appropriate and preserve ICIP?

The Australian Human Rights Commission in its 2021 Technology and Human Rights report recommended new technologies are built with deep consideration for human rights considerations – a human rights by design approach. This should also be a definitive goal for AI development and deployment. This would innately ensure inclusion of First Nations knowledge, perspectives and governance needs is done with the deep and genuine engagement needed to build the necessary trust to support First Nations communities' engagement with AI, as well as other community groups.

There are also clear risks that AI that draws on Indigenous knowledge or data without input – or sometimes even the awareness or appropriate permissions – from First Nations people runs the risk of extrapolating selective knowledge into other contexts, without due regard for the complexities of place-based cultural knowledge or deference for the appropriate protocols for Indigenous IP.

Science & Technology Australia recommendation 8

An additional guardrail should be added that specifies that any AI development or deployment must involve human rights by design principles that will lead to deep and genuine engagement with First Nations people for any AI that includes Indigenous IP or affects First Nations Australians or communities.

10. Do the proposed mandatory guardrails distribute responsibility across the AI supply chain and throughout the AI lifecycle appropriately? For example, are the requirements assigned to developers and deployers appropriate?

The mandatory guardrails for high-risk AI have the right foundations for supporting development, deployment and use. They acknowledge the need for assessment, monitoring and human oversight of operations. However, they could be further strengthened to create a cohesive, human-centric approach and support trust in AI.

For example, guardrails 1, 2, 3 and 8 set up accountability, risk management, governance and transparency processes, could be strengthened accordingly:

- Each developer and deployer within a supply chain should incorporate the previous year's assessment from the second year onwards to contribute to longitudinal risk management and accountability processes.
- Each developer and deployer along the supply chain should share their own risk assessment with other developers and deployers in the supply chain to generate understanding of the risk profile inherent in their product's or service's overall outcome.
- High-risk AI developed by large conglomerates should have an additional responsibility to respond to government, independent/commissioned reports and assessments.

Science & Technology Australia recommendation 9

To further strengthen accountability and transparency across the AI supply chain, the guardrails should be amended to require AI developers and deployers to develop and share longitudinal risk assessments.

11. Are the proposed mandatory guardrails sufficient to address the risks of GPAI? How could we adapt the guardrails for different GPAI models, for example low-risk and high-risk GPAI models?

GPAI does pose significant potential risks, which must be carefully managed. Experts broadly agree we need to better understand GPAI, which will require significant research investment to ensure Australia has the capability to develop and understand GPAI and can implement the highest standards in quality and safety for our GPAI use.

Science & Technology Australia recommendation 10

The Australian Government should invest in Australia's AI research capability to enable deep understanding and inform high-quality assessment and regulation of GPAI to ensure its safety and efficacy from development to commercialisation stages.

12. Do you have suggestions for reducing the regulatory burden on small-to-medium sized businesses applying guardrails?

The best way to minimise regulatory burden is to have a clear and consistent approach to implementing any regulatory regime. Specifying specific use cases for high-risk AI applications will also make it easier for businesses to navigate the regulatory requirements as it will remove grey areas and uncertainty.

The introduction of an appropriately resourced and responsive AI agency that could be responsible for overseeing Australia's AI effort/capability – from research to regulatory control – would provide a single point of contact for businesses. Developing such an agency would support Australia's sovereign AI capability and at the same time provide certainty, deep capability, and expertise in AI to support businesses navigating AI deployment and compliance.

Additionally, the government could consider implementing a tiered regulatory system according to system complexity (measured by required compute power) and/or business size and reach. Smaller businesses or those deploying smaller scale AI systems could be required to comply with a sub-set of guardrails, namely those ensuring appropriate human oversight over the AI system's decisions or outcomes.

Science & Technology Australia recommendation 11

To minimise the regulatory burden on small- to medium-sized businesses, the Australian Government should establish an Australian AI Agency to provide a single point of contact to support businesses navigate AI regulation requirements. The Government could also consider implementing a tiered approach to AI regulation and compliance, while still maintaining a mandatory focus on human-centric design and ensuring appropriate human oversight of any high-risk AI system.

Regulatory options to mandate guardrails

13. Which legislative option do you feel will best address the use of AI in high-risk settings? What opportunities should the government take into account in considering each approach?

AI will have – and is having – a whole-of-economy impact across society. As such, regulation to curtail risks should also take a clear and all-encompassing, whole-of-economy approach. This will ensure consistency in laws, support understanding of people's rights and obligations across all domains and regardless of the AI application.

The Government should develop an Australian AI Act that:

- takes a human-centric approach to developing all facets of AI regulation
- outlines the regulatory framework for high-risk AI, i.e. the principles to assess high-risk AI
- sets out the mandatory guardrails and requirements to address risks of high-risk AI
- establishes an Australian AI Agency to oversee Australia's AI engagement, including building a robust sovereign AI capability, drive industry investment and research in AI, support international engagement and cooperation, build Australia's AI workforce and manage AI regulation and safety
- is accompanied by a legislative instrument listing specific high-risk AI use cases.

An Australian AI Act would also provide the support needed to ensure consistent implementation of the proposed principles and mandatory guidelines. This would promote certainty and guidance for existing and future stakeholders including developers and deployers of high-risk AI.

Other options discussed in the consultation paper are piecemeal and may leave gaps that could reduce transparency, lead to exploitation or neglect of vulnerable groups and erode public trust.

Science & Technology Australia recommendation 12

The Australian Government should develop an Australian AI Act that enshrines the principles for assessing high-risk AI and sets out mandatory guardrails for high-risk applications, accompanied by a legislative instrument that lists specific high-risk use cases. The Act should also establish an Australian AI Agency to support and oversee Australia's AI sector, including regulation and compliance.

14. Are there any additional limitations of options outlined in this section which the Australian Government should consider?

The Australian Government should consider Australia's need to build a robust AI workforce, without which we will lack the expertise to properly understand or manage AI. CSIRO estimates that we will need up to 161,000 AI specialists by 2030. The robotics sector, closely aligned with AI, is also considered to be at critical risk of shortage. Our domestic capacity to build a deeply skilled AI workforce is narrowing with less than 10% of year 12 students studying advanced maths, a key prerequisite for AI, in 2020. Less than 1% of Australian university courses now require any higher maths. This sends a signal to students that maths isn't required, and students must spend significant time catching up to develop the required capability. While the migrant workforce can fill some gaps, Australia will be in increased competition with other countries for their expertise and we will need to stand out as an attractive destination.

Delivering a well-qualified and experienced AI workforce for both product and service developers as well as governance and accountability support will ensure Australia's security and mitigate risk of threats. An AI workforce is expected to have a range of occupations for its development including:

- Big data engineer
- Robotics engineer
- Natural Language Processing Engineer
- Research scientist
- Software engineer
- Business intelligence developer
- Machine learning engineer
- Data scientist[1]

No single one occupation will have the expertise needed on its own, so Australia must ensure a robust and comprehensive workforce covering all aspects of AI development and deployment.

The Australian Human Rights Commission in its 2021 Technology and Human Rights report, recommended human rights considerations are built into all new technology – a human rights by design approach. However, it noted that there were gaps in education and training—especially in STEM areas. A human rights by design approach should be a part of accreditation, ongoing professional development, training and capacity building for the entire AI workforce.

[1] **Source:** TAFE NSW Institute of Applied Technology - Introduction to Artificial Intelligence (AI) (Microskill course)

Science & Technology Australia recommendation 13

The Australian Government should consider programs and other ways to support the development of a deeply skilled AI workforce, including:

- targeted measures to improve advanced maths participation levels
- ways to incorporate a human rights by design approach to all AI development and deployment

15. Which regulatory option/s will best ensure that guardrails for high-risk AI can adapt and respond to step-changes in technology?

As outlined above, a two-pronged approach of an Australian AI Act that enshrines the principles to assess high-risk AI applications, accompanied by a legislative instrument to specify high-risk use cases will be the best way to address and respond to step-changes in technology.

16. Where do you see the greatest risks of gaps or inconsistencies with Australia's existing laws for the development and deployment of AI? Which regulatory option best addresses this, and why?

As noted above, the biggest risk to Australia's future prosperity lies in a failure to develop a deep sovereign AI capability alongside – and indeed, to inform – our AI regulatory environment.

Additionally, Australia's AI regulatory regime must also be implemented with clear understanding of the existing compliance burden on the industry and research sector. A risk-based approach to balancing the need for strong and consistent regulation with the flexibility and licences or freedom researchers and businesses need to do their best work must be implemented.

Please do not hesitate to contact Science & Technology Australia if we can be of any further assistance.

Professor Sharath Sriram	Ryan Winn
President	Chief Executive Officer
Science & Technology Australia	Science & Technology Australia