Response to proposed mandatory guardrails for AI in high-risk settings

On September 5th the Australian government released **voluntary AI safety standards**[1] for all organisations deploying or developing AI in Australia and a proposal for **mandatory safety standards for AI in "high-risk" settings**.[2]

This document contains the University's institutional response to the governments call for submissions in response to the proposal. It is based on our internal institutional experiences with AI governance to date. We recognise and respect the diversity of views within our broader community, many of whom continue to join this important international dialogue, and encourage them to make their own submissions.

Questions for consultation

1. *Do the proposed principles adequately capture high-risk AI? Are there any principles we should add or remove? Please identify any:*

    a. *low-risk use cases that are unintentionally captured*

Our response:

Principles (a) to (e) capture genuinely high-risk themes. However, principle (f) does not provide sufficient clarity to guide organisations on what meets the threshold of "high risk", meaning these would be hard to implement in practice. Example case studies would not be sufficient for this.

We recommend:

1. More detail on ways to rate these factors to provide practical clarity and market consistency. For example, for impacts are they:
    a. short term or long term;
    b. reversible, difficult to reverse or irreversible; and/or
    c. localised or systematic.

    For severity:

    d. the number and type of people impacted;
    e. the impacts on vulnerable groups; and /or
    f. the scope of release such as internal vs public.
2. Consistent with common risk assessment practice, that likelihood is factored into (f) as well as extent/severity.
3. Clarity on whether the risk assessment is "inherent" or made after risk mitigation measures are put in place by the developer or deployer.
4. Provision of standardised assessment tools to measure or rate factors of severity, extent, and likelihood of exceeding the risk threshold.
5. Provision of an evolving list of examples of high-risk applications to provide both specificity and flexibility given the rapidly evolving generative AI landscape.

Classifying all AI systems that use GPAI models as "high-risk" by default will unintentionally capture some low-risk use cases. Are these all intended to be captured, or only the GPAI models themselves? The University of Sydney has multiple GPAI chatbots and agents that we would not consider high risk enough to require legislative intervention – like other AI tools, risk depends on

[1] https://www.industry.gov.au/publications/voluntary-ai-safety-standard

[2] https://consult.industry.gov.au/ai-mandatory-guardrails

impact, severity and likelihood. For example, our policy and procedure chatbot could be high risk if released 'out of the box' to students. However, we would not consider it high risk to release it to select staff (such as policy teams) who can use it as a productivity tool and understand its limitations.

As technology and societal norms evolve, ongoing assessment will be needed to ensure low-risk applications such as educational or administrative uses aren't unintentionally escalated to high-risk classifications (particularly if GPAI model-based AI systems are automatically classified as high-risk).

b. *categories of uses that should be treated separately, such as uses for defence or national security purposes.*

Our response:

The question of sovereignty in AI development and deployment is critical. Australia's overwhelming reliance on: foreign-developed AI models; foreign SaaS providers selling AI systems; foreign companies providing the cloud compute necessary to develop or deploy AI systems; and foreign companies developing computer hardware - all present significant risks - not only in terms of national security but also for intellectual property, data protection, and economic independence. A stronger focus on encouraging the development of AI tools and expertise locally, while creating safeguards for foreign-developed AI models, is necessary to mitigate Australia's risk exposure in these areas.

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

Our response:

Consultation with Aboriginal and Torres Strait Islander leaders will be critical to developing AI systems that respect First Nations cultural protocols and are attuned to community-specific needs.

The University has had insufficient time to consult with our Aboriginal and Torres Strait Islander academic and professional staff leaders on this question. Our leaders have recently been giving these and related matters considerable thought to help strengthen the University's policy framework. We would welcome the opportunity to introduce the Department to our Aboriginal and Torres Strait Islander leaders for a discussion about these critically important issues.

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?*

a. *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?*

Our response:

We welcome uniform national guidance on AI risk and governance.

The principles-based approach is preferable as it offers the necessary flexibility to adapt to emerging technologies and use cases. A list-based approach, while providing concrete examples, could become outdated quickly in the context of AI's rapid evolution. Emerging applications, such as AI in autonomous decision-making or new forms of digital manipulation, may evade scrutiny if a rigid list is in place.

The example list provided also sets the bar significantly lower than the principles. For example, "evaluating learning outcomes" and "training" of staff are likely to become common and normalised use cases over time. These would not necessarily be high risk unless they also met the principles

b. *If you prefer a principles-based approach, what should we address in guidance to give the greatest clarity?*

Our response:

General: It should be clearer that a primary function of the 10 guardrails is to reduce the identified human risks identified by the proposed principles – not just business risks.

General: How will the guardrails interact with Australia's AI Ethics Principles? Will these remain voluntary for high-risk cases and/or do they need to be updated? We think it is important to take both an ethical and risk-based approach to development and deployment of AI.

Principle (a): the phrase "without justification" could be open to interpretation. It may be helpful to provide clear guidance on what constitutes a justifiable impact on rights (e.g. protecting public health, national security, etc.)

Principle (c): the phrase "similarly significant effects" is unclear.

Principle (e): the scope of "adverse impacts to the broader Australian economy, society, environment, and rule of law" is very wide.

Principle (f): while this principle provides for the severity and extent of impacts, it should also factor in likelihood. Guidance on how to measure or rate these factors is also required to provide practical clarity and market consistency. See 1(a) above. This would be particularly helpful for SMEs.

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?*

Our response:

Regardless of whether there are existing banned use cases, we support legislation that creates this category so emerging use cases can be added quickly if needed.

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

Our response:

Yes - as they focus on human impacts, they should be flexible enough to adapt.

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

Our response:

No. Applying mandatory guardrails across all GPAI models and any AI systems using them would be overly restrictive.

Development (through local modification of international models) and deployment of GPAI is already reasonably common, and will only be more so by the time legislation is passed. The proposal to class all GPAI tools as "high risk" would likely stifle local innovation, reducing the ability of Australia to benefit from GPAI.

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?*

Our response:

The same principles-based approach to risk assessment is broad enough to cover GPAI, at least for now.

Case studies involving generative AI chat bots and agents that incorporate third party GPAI, not just narrow AI, would be welcomed.

While technical metrics like FLOPS can be useful for certain evaluations, the complexity of AI systems often goes beyond computational power. High-risk indicators should include factors like the scope of impact, the potential for emergent behaviours, the contexts in which the AI is deployed, and societal implications.

Scientific panels and expert advisory groups should continually reassess these indicators as technology evolves.

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?*

Our response:

The guardrails as they are written, cover the key aspects of risk mitigation that would be needed to manage the use of AI in high-risk settings.

Of course, the effectiveness of the guardrails will also be impacted by:

- An engaged regulator who provides practical assistance to enable compliant use – we have already seen the benefits of having the NAIC.
- Magnitude and enforcement of penalties, with learnings from privacy regulation.

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

Our response:

See Question 2 above.

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?*

Our response:

Yes.

A supply chain-wide approach ensures that all stakeholders share responsibility for mitigating AI risks, reducing reliance on a single point of control or enforcement. Making deployers of high-risk systems equally responsible for compliance with the guardrails ensures they use their buying power accordingly.

Requiring transparency through the supply chain is also welcomed. Implementation of traceability mechanisms that document and verify compliance at each stage of the AI supply chain could enhance accountability.

More work needs to be done to fully consider the long-term continued reliance on foreign companies, regarding both AI system development and physical compute infrastructure.

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?*

Our response:

GPAI models can be applied in both low-risk and high-risk scenarios, making it more pragmatic to evaluate the context in which they are used. A flexible, principles-based assessment would allow for a more nuanced approach to defining the risk level of AI systems incorporating GPAI models.

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

Our response:

Automated systems for performing risk assessments and suggesting next steps, developed by the Australian Government and hosted securely on local compute infrastructure. Government-developed tools that help SMEs navigate regulatory requirements will be crucial, together with an engaged regulator. These systems should be easy to use, accessible, and tailored to different levels of AI expertise, with regular updates to reflect new risks and compliance measures.

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?*

Our response:

Option 3 presents the best solution for addressing the unique challenges posed by AI.

A dedicated AI Act with an appropriately engaged regulator can provide the needed clarity, consistency, and enforcement mechanisms to help Australia take advantage of AI effectively and safely.

This signals a clear position in a market where it has been our observation that many (even large) businesses are confused about what they can and cannot do, stifling innovation. It also gives us a national stance, which signals AI maturity in the international market.

In considering this approach, we agree that international interoperability is key given our relative size and computing power. The government should assess international frameworks, particularly the EU's AI Act and Canada's Artificial Intelligence and Data Act (AIDA). Both models provide valuable insights into building a regulatory foundation that protects public interest while fostering innovation. Drawing from these experiences, Australia could avoid common pitfalls, such as regulatory ambiguity or excessive red tape, while tailoring the approach to local needs and concerns, including national security and digital sovereignty.

The Australian Government should also prioritise the formation of a specialised, independent AI regulatory body that incorporates diverse expertise - spanning law, ethics, technology and public policy - to guide enforcement and monitor compliance.

The Voluntary AI Safety Standard is a critical piece. In the open market, this will give business consumers a clear framework to point to when selecting products for purchase, regardless of risk rating. We are grateful for the guidance provided to us as both a developer and deployer of this technology.

In time we expect that B2B consumers (and to a lesser extent B2C) will ask for third party certification of compliance with either the mandatory or voluntary guardrails as part of purchase/use, regardless of risk rating.

Careful thought should be given to regulation of "conformity assessment" providers to ensure quality and independence, without undue red tape, to balance innovation and governance. This may drive behaviour in unintended ways.

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

Our response:

In addition to the limitations already outlined in the sources, two significant concerns that the Australian Government should consider when evaluating its options for regulating AI are:

Regulatory Lag: The fast pace of AI development presents a substantial challenge to regulators. By the time legislation is implemented, technological advancements may already render certain provisions obsolete. This is particularly problematic if regulations rely on overly prescriptive or list-based approaches, which are inherently limited in responding to evolving AI capabilities and applications. The emergence of highly autonomous, agentic AI systems will raise new questions about accountability and liability, especially in areas where traditional legal frameworks may struggle to assign responsibility (e.g., systems acting without direct human intervention).

Additionally, AI models can develop emergent capabilities—unintended functionalities that were not initially programmed. While the scope of emergent effects remains debated, especially as it is often driven by commercial AI narratives, the potential for unintended behaviour highlights the need for future-proof regulations that remain flexible and adaptable.

Regulatory Capture: Another critical risk is the possibility of regulatory capture, where the interests of dominant technology companies unduly influence regulatory bodies. Given that Australia's AI ecosystem is heavily dependent on foreign technology, particularly from the US, there is a real risk that regulations could be skewed to benefit these large corporations. This could manifest in weaker guardrails, creating loopholes that allow for greater risks to be passed on to society while companies prioritise profit over safety and ethical considerations.

To mitigate this, it is essential to establish robust transparency and accountability mechanisms. Independent oversight bodies should be mandated to provide unbiased evaluations of AI technologies, and broad, meaningful stakeholder engagement must be a priority to ensure a diverse range of perspectives informs the regulatory process. This is particularly important in ensuring that Australia's approach to AI regulation supports sovereignty and economic resilience, reducing reliance on foreign companies and infrastructure.

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

Our response:

Option 3. However, this needs to include an engaged regulator who can provide guidance and support.

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?*

[We did not respond to this question]

General Response

Our response

We welcome the government's proposal to establish mandatory guardrails for AI in high-risk settings and appreciate the opportunity to provide feedback.

Our response:

- emphasises the need for a flexible, principles-based approach that can adapt to rapidly changing technology while providing key guidance for organisations on how to effectively and responsibly govern and implement AI;
- supports the creation of a dedicated AI Act while emphasising the need for an active and engaged regulator as the best path forward for Australia to safely and responsibly benefit from the opportunities made possible with AI;
- supports an internationally interoperable approach, especially given Australia's heavy dependence on foreign AI technologies;
- supports a strong focus on encouraging the development of AI tools and expertise locally, while establishing and enforcing safeguards for foreign-developed AI models;
- supports shared responsibility and transparency through the supply chain;
- seeks further guidance on what constitutes “high risk” with regard to impact, severity and likelihood of harms to humans, ideally across a scale of options; and
- challenges the proposal that all use of General Purpose AI (GPAI) is high risk, advocating for the same principles-based approach to be applied to GPAI and other emerging AI.