Swinburne University of Technology

Safe and responsible AI in Australia: Proposals paper for introducing mandatory guardrails for AI in high-risk settings

Authors:
Anthony McCosker, Natania Locke, Julia Tomassetti, Helen Bird, Yong-Bin Kang, Cesar Albarran-Torres, Bao Quoc Vo, Rowena Ulbrick, Ant Sowards and Virginia Kilborn.

Summary

Swinburne University of Technology (Swinburne) welcomes the opportunity to provide input into the Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings. As an innovative research and educational institution, Swinburne has a strong interest in supporting fundamental research into the use and impact of AI on society, and building the knowledge and learning outcomes needed to ensure Australia develops sovereign AI capability and talent in its graduates.

We firmly agree with the authors of the Proposals Paper that AI *"has the potential to increase productivity, address skill shortages, facilitate innovation and enable more tailored and accessible services, including in healthcare*." [p. 5] However, we also recognise the risks and harms inherent in some AI designs, and in many applications of AI including within educational environments.

In this submission we respond to Questions for which we have expertise and existing research interests. We submit the following specific points and recommendations:

- Australia has a relatively unique opportunity to lead in providing protections to First Nations people from harms through high-risk applications of AI. The broad capabilities and potential for harmful impact of GPAI warrants separate consideration and inclusion in the high-risk category for imposing mandatory guardrails to mitigate risks.
- We recommend including stakeholder engagement in the Mandatory Guardrails in line with the 10 guardrails for the Voluntary AI Safety Standards, Guardrail 10, in line with a *stakeholder first* approach to responsible AI development and deployment.
- Reducing regulatory burden in applying the guardrails must be supported by a greater focus on training, education and resourcing via a centralised regulator or expansion of the remit of an existing Office such as the Office of the Australian Information Commissioner. We welcome the consideration of support for AI capability through existing programs such as AI Adopt Program, and investments through the National Reconstruction Fund, but emphasise that more should be done to ensure education, training and capability development as a core component of responsible AI.

- We agree that Option 3 put forward in the consultation document would best suit the Australian legislative environment.

Detailed reasons and explanations are presented below.

Question group 1: Defining high-risk AI

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

A principle-based approach has greater capacity to accommodate new AI model developments and AI systems design and application, and more adequately covers risks and harms in the development and deployment of General Purpose AI (GPAI). However, as the consultation paper stands, the principles present more of a set of issues and potential harms for consideration. We suggest that the Proposed Principles be aligned with Australia's existing AI Ethics Principles, which already identify and define the key areas of risk and the associated goals for risk management in the development and deployment of AI.

We agree that a principles rather than a list approach be used to guide application of the mandatory guardrails for high-risk cases. A principle-based approach can be helpfully augmented with a dynamic and updatable list to illustrate domain specific applications and use cases as they emerge. However, such a list should not determine the definition of high-risk AI or limit the application of AI regulation.

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

We recommend adapting the emerging principles and practices for Indigenous Data Sovereignty as a basis for developing principles as a strengths-based approach to empowering First Nations People, communities and Country in the use of AI. The proposed guardrails and regulation seek to extend protections to cultural rights, with implications for the protection of First Nations intellectual property, cultural protocols and Indigenous data sovereignty.

There have been many studies revealing the inherent risks of bias in AI models impacting minority and already disadvantaged or marginalised groups including First Nations Australians. This includes the potential harms caused by, for instance, use of AI systems in predictive policing, surveillance and profiling, given the substantial over representation of First Nations people in Australia's criminal justice system. Instances of cultural harms can be seen in the training of GPAI models on First Nations artworks without permission or consent, enabling cultural appropriation and misuse of First Nations cultural expression. There is not a clear enough indication of how Intellectual Property protection will occur within the mandatory guardrails and principles approach to ensure that Indigenous art is not mis-appropriated as it is incorporated into image-based generative AI (GPAI) systems.

Australia has a relatively unique opportunity to lead in providing protections to First Nations people from harms through high-risk applications of AI. Many of the harms to "groups of individuals and collective rights of cultural groups" have their origin in the establishment and use or misuse of datasets that inform or underpin AI systems.

Question 3. Do the proposed principles, supported by examples, give enough clarity and certainty on high-risk AI settings and high-risk AI models?

Australia should include some high-risk uses of AI in the employment domain that are not in the EU and Canadian lists. First, it should include the use of AI in equipment that could affect someone's health and safety in the workplace, for instance, self-driving forklifts or robotic arms. (Preferably, the list of high-risk uses should include the use of AI in any equipment or device that could affect an individual's health and safety, not just in the employment domain). Second, the use of AI to direct work should be considered a high-risk case. For instance, the use of AI by platform companies or warehouses to direct the speed of work poses physical and psychological risks to workers and others. Similarly, an AI system that requires or encourages food delivery couriers to engage in hazardous driving poses a risk to the couriers and other road users.

Question 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?

We note that the consultation document does not propose a list of prohibited uses of AI technology, like what has been included in Article 1 of the European Union AI Act. The EU list seems to include abuses of AI technology that are by their nature contrary to human rights and cannot be justified by public interest. Australian legislation should adopt a similar list and ensure that these practices are subject to criminal sanction. For instance:

Manipulation by exploiting vulnerability: e.g., gambling, financial scams, drug and alcohol addiction, sexual exploitation and image-based abuse, mental health and psychological exploitation. Harmful manipulation of imagery includes nonconsensual intimate imagery or "deepfake porn", in which the image of a real person is used to make its likeness appear in AI-generated images or videos.[1]

Industries that rely on risky consumer behaviour, such as gambling and sports betting, alcohol and other drug consumption, use AI to tailor personalised experiences that can increase the chances of addiction in vulnerable individuals. Our research points to the difficulty in observing these practices as they take place through personalised feeds via

[1] McCosker, A. (2024). Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. *New Media & Society*, *26*(5), 2786-2803.

commercial social media platforms that do not have a strong track record of enforcing Australian regulation or platform community standards.[2]

Social scoring or ranking that would likely result in discrimination and social or economic exclusion.

Biometric identification without knowledge or consent including facial recognition or voice recognition in public places.

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

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

Question 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?

Several underlying features of GPAI models and systems position them as falling within the high-risk AI category.[3] GPAI models are designed to perform multiple tasks, unlike specific AI models that are usually tailored for a particular task. GPAI cannot be always defined as high-risk AI by default. However, the broad capabilities and potential for harmful impact of GPAI warrants separate consideration and inclusion in the high-risk category for imposing mandatory guardrails to mitigate risks.

GPAI may be considered high-risk, depending on several factors: 1) its intended application, 2) the scope of its autonomy, and 3) the potential consequences of its actions, and 4) privacy and ethical concerns:

Intended application: GPT models (GPT-3 and its successors) can perform various tasks from writing and summarising texts. However, these models can impact numerous sectors with high-risk consequences if misused. For example, a GPAI model can be used to predict legal

[2] Parker, C., Albarrán-Torres, C., Briggs, C., Burgess, J., Carah, N., Andrejevic, M., ... & Obeid, A. (2024). Addressing the accountability gap: gambling advertising and social media platform responsibilities. *Addiction Research & Theory*, *32*(4), 312-318.

[3] Md Meftahul Ferdaus, Mahdi Abdelguerfi, Elias Ioup, Kendall N. Niles, Ken Pathak, & Steven Sloan. (2024). Towards Trustworthy AI: A Review of Ethical and Robust Large Language Models, arXiv. https://doi.org/10.48550/arXiv.2407.13934.

Barua, S. (2024). Exploring Autonomous Agents through the Lens of Large Language Models: A Review. arXiv. https://doi.org/10.48550/arXiv.2404.04442

Judit Bayer (2024). The place of content ranking algorithms on the AI risk spectrum. Telecommunications Policy, 48(5), 102741.

Ross, A. AI and the expert; a blueprint for the ethical use of opaque AI. AI & Soc 39, 925–936 (2024). https://doi.org/10.1007/s00146-022-01564-2

outcomes that might inadvertently perpetuate historical biases present in training data. In the healthcare sector, if GPAI is used for diagnostic processes, it could misdiagnose patients if not accurately monitored or validated, thus leading to severe health consequences.

Scope of its autonomy: The autonomous nature of GPAI allows it to make decisions without human intervention, which can be risky in scenarios where nuanced human judgment is essential. For example, decision-making by GPAI in self-driving cars could lead to accidents if the AI misinterprets real-world data or faces scenarios not covered during training. In financial markets, trading algorithms based on GPAI could execute trades at high volumes and speeds, potentially leading to market manipulations if not properly regulated.

Potential consequences of its actions: The ability of GPAI to operate at scale can amplify any inherent biases or errors so that it can affect large populations extensively. For example, GPAI systems used to monitor and control online social media content can impact freedom of expression by over-censoring or failing to adequately filter harmful content.

Privacy and ethical concerns: When used in surveillance systems, GPAI can infringe on privacy rights if safeguards are not strictly enforced. Further, GPAI's capability to learn and adapt from new data can lead to unpredictable behavioural changes over time, potentially diverging from originally intended ethical guidelines. For example, a chatbot powered by GPAI that learns from user interactions could start generating harmful or inappropriate content if exposed to negative influences without proper filters. Also, GPAI-driven algorithms designed to personalize content on media platforms could spread misinformation by optimising for engagement over factual accuracy.

GPAI poses additional fundamental risks to Intellectual Property and Copyright. LLMs are trained on Australian data in non-Australian jurisdictions. There has been no mechanism to protect an Australian publisher or author in the development of GPAI in overseas jurisdictions outside of direct licencing deals. To address this issue in a systematic way, the overall AI regulation structure, and mandatory guardrails will have to be accompanied by mechanisms for enforcement and associated penalties to ensure accountability across the AI supply chain.

Question group 2: The ability of the mandatory guardrails to ensure testing, transparency and accountability of AI

Question 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?

While we welcome Guardrail 1 establishing accountability processes, the proposed guardrails as a whole need to directly ensure accountability by the whole of the organisations or entities involved in the development and deployment of AI in high-risk settings, not just those directly involved in the development and deployment of AI within those entities. A focus on

developers, deployers and end users within entities evidences a 'bottom up' approach. We suggest accountability should also be viewed as a systemic responsibility for the whole of the organisations or entities involved, from the those directly involved in its development or use, up through management and ultimately to the entities' boards of directors.

In assessing what guardrails might be appropriate in this context, it is appropriate to consider prudential statements published by the Australian Prudential Regulation Authority (APRA) on governance (CPS 510) and risk management (CPS 220). Specifically, on the requirement that boards of APRA regulated entities being required to provide an annual statutory declaration that their relevant entity has satisfied the regulator's prescribed requirements for risk management and governance.

We further refer you to the work of Elise Bant and others on systems intentionality (T*he Culpable Corporate Mind*, 2023) for more on the theory underlying these comments.[4]

Additional guardrails: We recommend including stakeholder engagement in the Mandatory Guardrails in line with the 10 guardrails for the Voluntary AI Safety Standards, Guardrail 10: "*Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness*." This guardrail is equally and perhaps more important for high-risk AI.

A *stakeholder first* approach to responsible AI development and deployment can support the scrutiny of risks and harms more effectively than internal or self-audits.[5] Stakeholder engagement presents an opportunity to boost the inclusiveness of AI systems through either their development phase or deployment, improving the likelihood of representation of those most likely to be affected or potentially harmed by the deployment of AI systems.[6] The absence of stakeholder engagement in the mandatory guardrails creates a substantial gap in the capacity to identify emergent risks and harms as they impact diverse people who use or engage with AI systems.

---

[4] Bant, E. (2023). Systems intentionality: Theory and practice. In Elise Bant (Ed), *The Culpable Corporate Mind*, Oxford: Bloomsbury Publishing, p. 184.

[5] Domínguez Figaredo, D., & Stoyanovich, J. (2023). Responsible AI literacy: A stakeholder-first approach. *Big Data & Society*, *10*(2), 20539517231219958. Bell, A., Nov, O., & Stoyanovich, J. (2023). Think about the stakeholders first! Toward an algorithmic transparency playbook for regulatory compliance. *Data & Policy*, *5*, e12.

[6] Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022, October). Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-8). McCosker, A., Yao, X., Albury, K., Maddox, A., Farmer, J., & Stoyanovich, J. (2022). Developing data capability with non-profit organisations using participatory methods. *Big Data & Society*, *9*(1), 20539517221099882.

Results of the study from UTS would also support engaging with workers as stakeholders in the decisions about the development, deployment, and use of AI systems in the workplace.[7] In the sectors studies in the report, workers were generally left out of discussions about AI in their workplaces and consequently did not trust the system and felt the use of AI created serious risks.

Civil society and not-for-profit organisations can play a leading role in establishing best practice in identifying emergent AI risks for communities and target groups and mitigating harms through their capacity to support or advocate for vulnerable and marginalised people and groups. Key stakeholders across parts of civil society can support the implementation of the guardrails, to ensure individual, group and cultural harms can be avoided, including financial services, community health, disability, aged care, employment services, family services, housing and homelessness services, First Nations organisations and consumer rights organisations.

Civil society organisations and not-for-profit organisations are well placed as key stakeholders in mitigating the harms of high-risk AI through their experience and expertise in dealing with sensitive data, maintaining privacy rights, autonomy and dignity for marginalised and disadvantaged groups in society. Our research recognises the urgent need for supporting data and AI governance practice across public services, the NFP sector and public health sector.[8]

Question 9. Inclusion of First Nations knowledge and cultural protocols to ensure AI systems are culturally appropriate and preserve ICIP.

Australia has a relatively unique opportunity to lead in providing protections to First Nations people from harms through high-risk applications of AI. The proposed guardrails and regulation seek to extend protections to cultural rights, with implications for the protection of First Nations intellectual property, cultural protocols and Indigenous data sovereignty.

[7] UTS Human Technology Institute (2024). 'Invisible Bystanders: How Australian workers experience the uptake of AI and automation', Sydney: UTS, https://www.uts.edu.au/sites/default/files/2024-05/EssentialResearch%2BUTS_Invisible_Bystanders_0524_D4.pdf

[8] Kang, Y.B., McCosker, A., Savic, M., Graham, T., Jayaraman, P.P. (2023) AI Governance in the Smart City: A case study of garbage truck mounted machine vision for roadside maintenance. Swinburne University of Technology, Melbourne, DOI: https://10.25916/a2fn-yb49.

McCosker, A., Yao, X., Albury, K., Maddox, A., Farmer, J., & Stoyanovich, J. (2022). Developing data capability with non-profit organisations using participatory methods. *Big Data & Society*, *9*(1), 20539517221099882.

McCosker A., Shaw F., Yao X., Albury K. (2022) A Data Capability Framework for the Not-for-Profit Sector. Swinburne University, Melbourne, DOI: https://doi.org/10.25916/h2s6-r178.

Albury, K. and Mannix, S. (2023) Digital and Data Capabilities for Sexual Health Policy and Practice: Stage One Report. July 2023. Melbourne: Swinburne University of Technology. https://doi.org/10.25916/782g-nk95

The National Indigenous Australian's Agency (NIAA) Framework for Governance of Indigenous Data supports monitoring and accountability in the use of Indigenous data, including information and knowledge in any format or medium, as part of a right to self-determination. Combining this with a stakeholder first approach to responsible AI development and deployment offers mechanisms for ensuring that First Nations people's rights are protected.

Stakeholder engagement for high-risk AI is essential to enable incorporation of Indigenous data sovereignty principles and practices, requiring further examination of pathways to ensure First Nations organisations and communities are considered key stakeholders in AI development and deployment that affects them directly.

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

We agree that SMEs, and in particular not-for-profit organisations will require support in applying both mandatory and voluntary guardrails. Whether in the application of mandatory or voluntary guardrails, there has not been time for building the relevant knowledge and expertise across sectors seeking to develop or deploy AI. With the availability of GPAI and its adaptability in applications like service sector chatbots, real world deployment is moving quickly. Much of this development and deployment will be in low-risk categories, but even in low-risk applications, applying the guardrails and building community trust in AI will require new investment in training, education and upskilling.

Our research with local government found that there is much work to do in implementing AI governance to support the implementation of AI systems among public services.[9] We supported a local Council in Victoria to develop an AI governance framework and action plan to ensure the ethical and safe implementation of its roadside maintenance monitoring program via 5G connected machine vision mounted rubbish collection vehicles. The work found that there was little guidance and a need to build capacity in local government and the public sector more broadly for effective AI governance. Local governments, like many small-to-medium enterprises need support improving data governance and policy frameworks that can adequately respond to rapid digital transformations.

Reducing regulatory burden in applying the guardrails must be supported by a greater focus on training, education and resourcing via a centralised regulator or expansion of the remit of an existing Office such as the Office of the Australian Information Commissioner. We welcome the consideration of support for AI capability through existing programs such as AI Adopt Program, and investments through the National Reconstruction Fund, but emphasise

[9] Kang, Y.B., McCosker, A., Savic, M., Graham, T., Jayaraman, P.P. (2023). AI Governance in the Smart City: A case study of garbage truck mounted machine vision for roadside maintenance. Swinburne University of Technology, Melbourne, https://10.25916/a2fn-yb49, https://apo.org.au/node/323811

that more should be done to ensure education, training and capability development as a core component of responsible AI.

Question group 3: Regulatory options

Question 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?

We agree that Option 3 put forward in the consultation document would best suit the Australian legislative environment.

This option is in line with the approaches in the European Union and in Canada. There are sound reasons why international harmonisation in AI regulation must as far as appropriate be adopted. This approach should reduce barriers to entry for new deployers, as they will be familiar with basic principles that apply regardless of jurisdiction. It will also encourage international investment, as the adoption of familiar principles will give comfort to investors that a similar measure of regulation applies in the jurisdiction of the business to be financed. This underlying theory is similar to the approach for securities regulation, where harmonisation is driven by the International Organisation for Securities Commissions. Potential applications of AI are not bound to jurisdiction or territory, which further drives the need for universal principles for safe and responsible use.

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

A major limitation of Options 1 and 2 is that regulatory monitoring and enforcement is left to distinct areas, rather than based in an over-arching regulator with specialised capacity, enforcement powers, and resources. This approach risks regulators adopting a 'silo' mentality to their monitoring and enforcement work and downplays the possibility of early identification of an economy wide AI risk or threat. This is a primary reason why we prefer Option 3.

History has shown that attempts to put down regulatory principles without the support of monitoring and enforcement powers typically lead to poorer adoption and compliance. A good example of this is in the terrain of sustainability reporting, where voluntary codes and reporting systems were first widely attempted. This led to cherry-picking by business of the codes that best supported their current practices. In some instances, business entities simply ignored the voluntary guidance, especially if they were not subject to securities exchange requirements. This had led to a global move to mandate sustainability reporting.

The employment domain offers an example of the limits of Australia's current regulatory system and of Options 1 and 2 to promote safe and responsible AI. As the Proposals Paper notes, Option 1, relies on existing domain-specific regulators and their enforcement powers. This is likely to limit the ability to implement pre-market guardrails where the domain-specific regime relies primarily on ex post enforcement, as is the case with employment regulation. As recognised, however, the use of AI systems in employment, for instance, in hiring and performance evaluation, poses risks of illegal discrimination. Anti-discrimination law generally does not impose the equivalent of 'pre-market' obligations in the workplace where there is a risk that new practices will have discriminatory effects: it includes little scope for the regulator to intervene before development or deployment. Thus, anti-discrimination law does not require employers to undertake risk assessments or perform testing before introducing practices that pose a risk of illegal discrimination. (The Sex Discrimination Act (SDA) does prohibit an employer from 'propos[ing]' to impose a discriminatory workplace condition or practice; however, in almost all cases, the employee takes legal action against the employer only after the imposition.) Further, most regulatory bodies responsible for enforcing anti-discrimination law, do not have the power to initiate civil proceedings.

Even where employment law imposes 'pre-market' duties, under Options 1 and 2 the effectiveness of the guardrails would likely be curtailed by the regulator's limited scope of authority and technical expertise. For example, the SDA was recently amended to include a positive (aka a 'pre-market') duty for employers to eliminate sex discrimination. Under Options 1 or 2, it is plausible that the Australian Human Rights Commission (AHRC) (the body responsible for enforcing the new duty) would determine that an employer, to fulfil the new duty, must carry out a risk assessment and tests before using an AI system to make important personnel decisions. However, as acknowledged in the Proposals Paper, it is often the developers who are responsible for the high-risk elements of an AI system. As a consequence, where the employer is a deployer rather than a developer, the regulatory scope of anti-discrimination law may not reach the party best placed to prevent discrimination (for instance, where the discriminatory effect of the AI system is a product of the model or source data). And, even if it could reach that party, given that AI systems, and particularly GPAI systems, are complex technical systems, the regulator may not have the technical capability to assess compliance with anti-discrimination law.

The capacity of domain-specific regulators is also likely to limit the effectiveness of the pre-market guardrails under Options 1 or 2. Workplace health and safety law requires employers to take a proactive approach to providing a safe workplace. And regulators have power to act before any injury occurs. Further, they tend to use a risk-based approach to allocating resources and seek to balance proactive and reactive measures in enforcing the law. Nonetheless, in practice, it appears uncommon for the regulator to intervene before an employer adopts a workplace practice or new equipment that poses a safety risk. (Only for

certain high-risk instances do regulators require pre-notifications (like asbestos removal). Thus, Option 3 would best protect workers where the employer adopts a high-risk AI system, for instance, AI-controlled forklifts, by requiring pre-market guardrails and placing the responsibility for enforcing them in a dedicated regulator.

Option 3 will offer the advantage of housing specialised skill for providing guidance and monitoring compliance with the guardrails. This could then be used in the enforcement action of a new regulator, as well as in support of the enforcement actions of other regulators. Some overlap may exist, but this can be governed during the legislative drafting process. For instance, cooperation agreements may be put in place between the new regulator and the Australian Securities and Investments Commission and the Australian Competition and Consumer Commission, so that duplication of enforcement does not occur. This system is well-established in Australia.

Option 3 will also complement the government's objective of building Australia's AI capabilities (Attachment B, no. 3 of Proposals Paper). This is critical to ensure that AI expertise and the trajectory of the development of AI are not monopolised by the private sector. The government's approach to AI should be informed by cautionary tales from other industries where expertise, technical capabilities, and resources were over-concentrated in the private sector. Consider, for example, the tobacco industry's success in delaying public knowledge of the causal connection between smoking and cancer due in part to their efforts to control and manipulate scientific research and funding.[10] The government's Rapid Response Information Report indicates that we may already have cause for concern here. It notes that the US is facing a 'critical undersupply' of graduates qualified in machine learning because tech sector recruitment from universities has left too few faculty able to supervise PhDs.[11]

Further, although we realise that holding developers and deployers accountable under product liability law is beyond the scope of the Proposals Paper, the government should consider the interaction between pre-market and post-market regulation. While we think the 'development risk' or 'state of the art' defence under the Australian Consumer Law should be eliminated or restricted for AI systems, assuming it is not, we note that leaving the private sector in control of AI development could make it easier for developers to rely this defence when their products cause harm: The developers could determine the state of scientific and technical knowledge (like the tobacco industry) but have little incentive to develop that knowledge in any manner that could slow down or jeopardize the marketability of their

[10] Hanson, J. D., & Kysar, D. A. (1998). Taking Behavioralism Seriously: Some Evidence of Market Manipulation. *Harvard Law Review*, *112*(7), 1420–1573.

[11] Bell, G., Burgess, J., Thomas, J., and Sadiq, S. (2023, March 24). Rapid Response Information Report: Generative AI - language models (LLMs) and multimodal foundation models (MFMs). Australian Council of Learned Academies.

products. Option 3 is preferable to ensure the government develops the expertise necessary to regulate AI both pre- and post-market.
There is also a risk of a lagging of legislative amendment in other industries if Option 1 or 2 is adopted, not to mention the cost and time involved in the legislative process. Options 3 guarantees a uniform approach.

Limitations include issues of interoperability and global alignment to govern foreign AI development. Alignment with North American and UK and EU approaches will be necessary to address the overseas development of systems. However, this could introduce potential misalignment with important trading partners and developers of AI systems and products including through alternative approaches being considered by BRICS countries including India and China, and ASEAN countries in Australia's region.

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

We recommend the inclusion of broad principles in the main legislation, supported by delegated legislation and/or regulatory guidance to respond to unforeseen instances where the market needs additional guidance in application.

It is virtually impossible to predict the specific uses and future development of AI. Principles-based regulation is therefore better suited to this environment. The principles may be supported by non-exhaustive lists of known examples, similar to what had been adopted in the comparative jurisdictions, but we recommend that such lists be included in accompanying regulations to enhance the ease with which new use cases may be included. This is similar to the approach in Canada.

Principles-based regulation should be supported by an open-door policy by the regulator, so that innovators may run their plans and risk mitigation strategy past the regulator before implementation. Regulatory sandboxes may a good option in this space. Resources should be allocated to the regulator in support of these functions. International cooperation between similar regulators may also be of assistance.

We note that the Canadian legislation does not apply to AI employed by government, unless developed by the private sector. Unless separate specific legislation is proposed for governments in Australia, there is no reason for such a blanket carve-out in the local setting. The government of Canada is subject to other regulatory safeguards in this area, which is not present in Australia.

Question 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?

Please also refer to our response under Question 14 in the context of this question.