Submission to the Department of Industry, Science and Resources Consultation on 'Safe and responsible AI in Australia'

August 2023

Authors: Professor Nicole Gillespie and Dr Steve Lockey, The University of Queensland
Contact: n.gillespie1@uq.edu.au

The DISR have called for public submissions to inform Government policy on the safe and responsible use of AI in Australia. The Discussion Paper highlights that public trust in AI technology is required to realise the economic and societal benefits of AI. It also outlines the Government's commendable intent to develop a clear set of regulatory and governance mechanisms to support and guide the trustworthy use of AI to protect Australians from harm and strengthen public trust in the use of AI.

This intent aligns directly with the recommendations identified in our recent research report examining public trust and attitudes towards AI in 17 countries, including Australia[1]. Using nationally representative samples, this research shows that only a third of Australians are willing to trust AI systems and less than half of Australians perceive the benefits of AI applications as outweighing the risks. The report more broadly highlights that the public have high expectations of the governance, regulation, and management of AI.

These findings underscore the importance of government and industry addressing the significant concerns the Australian people have with the use of AI, and the need to strengthen the regulation and governance of AI to augment 'warranted trust': trust that is founded on and calibrated with the responsible and trustworthy use of AI.

In this submission, we draw on this research together with insights from our systematic literature review on trust in AI[2] and case study research examining stakeholder trust in AI-enabled public services, to provide evidence-based responses and recommendations to the following question posed in the Consultation Discussion Paper: ***What initiatives or government action can increase public trust in AI deployment, and support responsible AI practices in Australia?***

We keep our recommendations and reporting of the research to a high level and would be pleased to brief DISR more fully on any aspect of the research or recommendations.

Recommendations:

1. ***Develop a clear, consistent, risk-based regulatory framework for governing AI that has effective independent oversight.***

    - Our research demonstrates that a strong predictor of people's trust in AI systems is the belief that there are adequate regulations, laws, and safeguards to make AI use safe.

    - The Australian public clearly expect a stronger set of regulatory mechanisms and laws to protect them against the perceived risks of AI use. Our research shows that 70% of Australians expect AI to be regulated, however the majority (65%) do not believe current regulations, laws and safeguards are sufficient to make AI use safe. This suggests that the regulatory and legislative frameworks supporting trustworthy AI are failing to keep pace with community expectations.

[1] Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). *Trust in Artificial Intelligence: A Global Study.* The University of Queensland and KPMG Australia. doi:10.14264/00d3c94

[2] Lockey, S., Gillespie, N., Holm, D., & Someh, I. A. (2021, January). A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, (pp. 5463-5472).

- Australians have a clear preference for AI to be regulated by government and existing regulators, or by an independent AI body, rather than by industry. Australians also expect industry to play a role through co-regulation. We recommend a range of hard and soft regulatory mechanisms be implemented, with hard regulatory mechanisms required for higher risk AI applications.

- Given the widespread deployment of AI, it is recommended that a risk-based approach to regulation is adopted. Implemented well, a risk-based approach enables regulatory and governance efforts to focus on mitigating risks and preventing harm from moderate and high risk AI applications. This enables regulation to be practical and effective while supporting innovation and rapid application of AI in low risk applications. A risk-based approach is appropriate given the vulnerabilities and risks associated with AI are application and context specific, and therefore require a proportionate and tailored response[3].
  Risk based approaches can promote the use of AI for beneficial, human-centred purposes by requiring moderate to high risk AI applications to demonstrate a clear beneficial purpose to people and society. Our research demonstrates that AI applications that have a tangible, beneficial impact on people and/or society are more trusted.

- To be effective and meet public expectations, it is recommended that AI regulation and governance frameworks be informed by and seek to uphold the Trustworthy AI principles proposed by the European Commission[4], which are inclusive of Australia's AI Ethics Framework[5].
  Our research demonstrates strong endorsement for these principles, with 95% of Australian's stating these principles - and the practices underlying them - are important for their trust in a range of common AI applications. These principles include:
  - *Accountability*: the entities accountable for the use and outcomes of AI systems are identified, with clear allocation of responsibility across the design, development, deployment, and post-deployment phases of the AI lifecycle
  - *Contestability*: people impacted by AI systems can contest and challenge outcomes via a fair and accessible process
  - *Fairness, non-discrimination, and diversity*: AI systems are inclusive and accessible and do not result in unfair discrimination of individuals or groups.
  - *Transparency and explainability*: there are transparent and responsible disclosures to enable people to understand when and how they are significantly impacted by AI, and when they are engaging with an AI agent.
  - *Data privacy and security*: people's right to privacy and data protection are upheld, and data is kept secure.
  - *Technical performance and robustness*: the performance, reliability and accuracy of AI output is assessed before and during deployment to ensure it operates as intended.
  - *Human agency and wellbeing*: AI systems are used to benefit individuals and society and support human autonomy, agency, and human rights.
  - *Risk and impact mitigation*: The potential risks and harm from AI systems are assessed and mitigated prior to and during its deployment, with appropriate human oversight of AI systems that significantly impact people.

- We recommend requiring human review and oversight of AI applications deemed moderate to high risk or where AI outcomes have significant consequences for people. Our survey and case study research suggests that people expect a human in the loop when AI systems are used for consequential decision making about people, with low support for automated

[3] Gillespie, N., Curtis, C., Bianchi, R., Akbari, A., and Fentener van Vlissingen, R. (2020). A*chieving Trustworthy AI: A Model for Trustworthy Artificial Intelligence*. KPMG and The University of Queensland Report. doi.org/10.14264/ca0819d

[4] AI HLEG. (2019). *Ethics Guidelines for Trustworthy AI*. European Commission. Retrieved from https:// ec.europa.eu/

[5] *Australia's AI Ethics Principles*. Retrieved from https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ ai-ethics-principles

decision making. In addition to a human in the loop, we further recommend that moderate to high risk applications require: a) AI impact assessments to be conducted and reviewed, b) transparent notification of the use of AI, c) accessible explanations of how the AI output is produced, and d) ongoing monitoring and reporting of the performance of the AI solution.

- Where appropriate and feasible, it is recommended that existing laws and regulatory frameworks be adapted to be fit-for-purpose for mitigating the potential risks and harms of AI. We note that that some novel applications of AI will require new laws and regulations to be developed.

- To be effective and efficient, Australia's AI regulatory framework should be consistent and aligned across all Australian jurisdictions, including federal, state and territory levels.

2. ***Support the alignment of Australia's regulatory and governance approaches with those in other jurisdictions and emerging international AI standards.***

   - Aligning Australia's approach with the emerging regulations and laws proposed by the EU AI Act and other jurisdictions is important to enable the use, export and import of AI-enabled goods and services across borders, and a clear and consistent approach to mitigating harm.

   - Our research indicates people view the risks of AI in a comparable way across the globe, with 73% of people reporting concerns about cybersecurity and privacy breaches, loss of jobs and deskilling, manipulation and harmful use, system failures, the erosion of human rights, and inaccurate or biased outcomes. Cybersecurity is the top-ranked concern in Australia and one of the top two concerns across all countries surveyed. Further, we find global endorsement for the principles of trustworthy AI with 95% or more of people in all countries stating that these principles are important for them to trust in AI applications.

   - These findings highlight the merit in supporting global collaborative approaches to governing and regulating AI, as well adopting international standards for mitigating AI risks, and supporting responsible AI.

3. ***Invest in communicating and educating the public, industry and government on the regulations and laws governing AI, and support industry and government to adhere to regulatory and governance frameworks.***

   - Efforts to strengthen the AI regulatory framework will only translate into enhanced public trust if a) the public are aware of the laws and regulations and understand how they help to safeguard and protect them, and b) industry and government adhere to the regulation. Hence, it is important to invest in public and industry education campaigns to raise awareness, understanding and adherence.

   - It will be particularly important to provide support and guidance to SMEs, nonprofits, and local government to facilitate their adherence with AI regulations and appropriate governance frameworks.

   - It is recommended that research be commissioned to periodically evaluate and report on public and industry awareness, understanding, adoption and effectiveness of AI regulatory and governance frameworks over time.

4. ***Invest in developing the capability and processes required to enable AI regulatory and governance frameworks to a) be regularly reviewed and updated to keep pace with the rapid evolution of AI and emerging technologies, and b) be enforced.***

   - Given the ongoing rapid evolution and novel application of AI technologies, it is recommended that the government invests in developing the capability and processes required to enable the rapid review and amendment of relevant laws, regulations, and governance frameworks over time. The effective mitigation of risk and prevention of harm from evolving AI technologies necessitates the efficient and rapid adaptation of regulation to ensure it is fit for purpose. This is also important to overcome the strong current public perception of ineffective regulation and regulatory lag noted above. AI regulatory sandboxes are one mechanism to enable this.

- To be effective, it is recommended that the government invest the required resources and capability to ensure the AI regulatory framework can be enforced in practice. Increasing the public's perception of a strong and effective regulatory framework is likely to augment the perceived trustworthiness of AI systems, increasing trust and willingness to adopt and rely on AI systems. Such increased trust, reliance and adoption will augment risks and the potential for harm over time if it is not accompanied by the ability to effectively enforce laws and regulations and impose sanctions against those that violate regulations.

- It is recommended that the regulatory framework be developed and regularly evaluated, reviewed, and adapted in consultation with industry, civil society, and experts to ensure it is practical, effective, and enforceable.

5. ***Establish an independent AI Commission***

   - The Australian Human Rights Commission have recommended the creation of an AI Commissioner that acts as an independent statutory authority to support and provide expert advice to government, regulators, and industry on the safe and responsible use and governance of AI.

   - We support this recommendation and note that such a commission could play a pivotal role in supporting recommendations 1-4 above. We also envisage an AI Commission could help raise awareness of responsible AI and regulatory and governance expectations and requirements and support the development of aligned policy across levels of government.

6. ***Develop and implement a consistent AI Assurance Framework across all levels of government that integrates leading practices in AI Impact Assessment, AI ethical review, governance and risk management, procurement, and assurance mechanisms to support the responsible and trusted use of AI.***

   - Our research shows that 75% of people are more willing to trust AI applications when assurance mechanisms are in place that demonstrate the responsible deployment of AI systems. These mechanisms include assurances that the accuracy and reliability of AI systems are monitored, AI systems are reviewed by an independent AI ethics board and comply with international standards for explainability and transparency, and organizations using AI are bound by an AI code of ethics. These mechanisms provide assurances that safeguards are in place, reducing uncertainty and supporting trust.

   - Our case study research suggests that the NSW AI Assurance Framework has been effective in supporting, guiding, and incentivising government agencies to design, develop, and deploy AI in a trustworthy way. This assurance framework serves as a model that can be adapted and updated for wider application across government, and potentially industry. This framework incorporates risk and impact assessments of AI projects, documentation of decision making, and review by an independent AI review committee where required, with part of its effectiveness tied to its mandatory nature to obtain funding. Our research suggests that it is more difficult to gain support for AI projects in government agencies that do not have an established and accepted AI governance framework.

   - Our case study research indicates many government agencies are at an early stage of maturity in understanding AI and its responsible governance. We recommend investment in building a foundation of data and AI capability and expertise in government departments as a first step, as well as adopting clear governance and accountability processes such as those articulated by the NSW AI Assurance framework.

- We recommend federal and state governments create a consistent approach to AI procurement and vendor management processes that supports and incentivises the trustworthy procurement and deployment of AI and sets minimum required standards. Such a procurement approach will need to clarify accountabilities and responsibilities and require documented evidence and audit trails on the testing, performance and ongoing monitoring of AI solutions deemed to be of moderate to high risk.

- In situations when AI systems are being used to enable the delivery of public services, we recommend the use of codesign and consultation processes with key stakeholders, as well as staged deployment through piloting, given evidence that these processes support stakeholder trust and trustworthy AI deployment.

- The effective implementation of an AI Assurance Framework across government would help position the government as a leader in the trustworthy deployment and governance of AI, supporting its credibility to regulate and govern AI use by industry.

7. ***Invest in upskilling people to understand and responsibly use and engage in AI and digital technologies to support trust, adoption, and equity of access.***

- Our research shows that people are more likely to trust AI applications when they understand AI and when and how AI it is commonly used and feel competent in using digital tools in their everyday life. Yet, most Australians have low understanding of AI. 65% of Australians report they want to learn more about AI, indicating an appetite for a public AI education program.

- We recommend implementing public and customer education programs to augment AI understanding and capability in using digital technologies to support trust and equity of access to digital services. Such an education program can augment people's understanding of the benefits and risks of AI to enable them to make informed decisions and avoid inappropriate use.
  As more products and services become fully digitized, it is important to support people's efficacy with, and access to, online tools. Our research suggests such programs should particularly support older generations and people without a university education, given these groups report lower levels of understanding and digital capability compared to younger generations and those with a university degree.

- We recommend government, universities and industry collaborate to uplift public and consumer literacy and understanding of AI, as well as appropriate data and technology use.

8. ***We recommend transparency in informing people when they are interacting with an AI agent and not a human, and when AI is involved in consequential decision making about individuals.***

- Transparency will be most valuable to mitigating potential AI risks and improving public trust and confidence in AI in situations where:
  a) the AI is anthropomorphic (human-like), and people may not know they are interacting with an AI agent rather than a human. This is important to prevent manipulation and harmful use.
  b) the AI is either making or significantly informing decisions of consequence about people, and hence may be negatively impacted if the AI output is inaccurate or fails.