UNSW AI Institute
ai.director@unsw.edu.au

25 July 2023

Department of Industry, Science and Resources
Industry House
10 Binara Street
Canberra ACT 2600

Dear Sir/Madam,

Thank you for this opportunity to respond to the public consultation on the Discussion Paper on "Safe and responsible AI in Australia". This submission contains the collective advice of UNSW's flagship research institute in artificial intelligence. The central goal of UNSW.ai is the responsible development and deployment of AI. We therefore welcome the government's attention to this issue.

For context, UNSW's new AI Institute brings together over 300 world-class researchers, across UNSW Sydney and Canberra, working on fundamental AI problems as well as applying AI techniques to their respective domains. Whilst addressing the legal, societal, and economic implications of AI. UNSW.ai is the largest research institute in Australia focused on AI. Members of the institute have previously provided advice on AI to government departments, academies, parliamentary inquiries, company boards, startups, charities, and other non-governmental organisations.

We especially thank the following researchers for their comments and feedback: Beena Ahmed, Kate Carruthers, Matt Garratt, Dong Gong, Fei Huang, Aditya Joshi, Rachida Ouysse, Asanka Perera, Imran Razzak, Flora Salim, Eduardo Benitez Sandoval, Sebastian Sequoiah-Grayson, Vidhyasaharan Sethu, and Yang Song. Please also refer to the independent report by UNSW Allens Hub for Technology, Law and Innovation that takes a deeper dive into legal aspects.

Please do not hesitate to contact Dr Haris Aziz at ai.director@unsw.edu.au should you wish to discuss any issue raised in this submission.

Yours faithfully,

Dr Haris Aziz Acting Director UNSW.ai	Professor Toby Walsh Chief Scientist UNSW.ai

UNSW AI Institute Submission: Safe and responsible AI in Australia

Overview

1.0. Although some of the challenges of ensuring that AI is deployed responsibly may be addressed by enforcing existing generic rules (such as those around privacy) and domain specific legislation (such as that around financial services or medical devices), the majority of the challenges and risks are not yet addressed by the existing ethical and regulatory frameworks. We must ensure that sufficient AI expertise sits within the bodies tasked with regulatory responsibilities, and that there are sufficient resources available to the relevant agencies tasked with undertaking this work.

1.1. Artificial intelligence does throw up several unique challenges. One of the most important is the speed and scale with which AI technologies can be deployed. A consequence of this is that there is some urgency for government to respond to these challenges.

1.2. Artificial intelligence is a general-purpose technology that will become a pervasive part of our homes, offices, and factories. It would be appropriate to oversee some of these widespread applications using specific AI focused regulation.

1.3. AI does not respect geographical borders. Australia's approach needs therefore to be compatible with the approaches of other regions. To this end, the forthcoming EU AI Act is an obvious reference point for Australia as we formulate our own position. A risk-based approach such as found with the EU AI Act has much to offer.

1.4. At a minimum, we need to develop and pursue design guidelines, and a code of ethics and practice that apply to AI development in Australia in a similar format as the code of ethics at IEEE or ACM.

1.5. Australia needs to cooperate and keep pace with the global leaders in the AI regulatory eco-system. 2022 saw the general approach of the EU AI Act adopted in December, the publishing in October of the United States AI Bill of Rights, and in July of the United Kingdom's AI Regulation Policy Paper. In March, China saw the enforcement of its Algorithmic Recommendation Management Provisions.

Key Recommendations

We make the following six recommendations:

2.0. Investment
AI will not be deployed responsibly without significant government investment. The Australian Government needs to invest in basic AI research (e.g., to provide technical solutions to fundamental issues like measuring and achieving fairness, as well as to identify new and emerging harms), translation (e.g., to enable Small and Medium Enterprises (SMEs) to adopt AI responsibly) and in regulatory bodies overseeing the responsible deployment of AI.

2.1. Regulation
Much of the harm that AI could bring can be addressed within existing regulatory frameworks. There are, however, some new threats that cross sectors, such as the use of AI in biometric identification or in employment. These could be regulated with some AI-specific regulatory frameworks along the lines of that being proposed in the EU, but this must be done with careful adaptation to the local contexts, incorporating the diversity of values on the ground.

2.2. Procurement
The government has considerable power to set standards for the responsible deployment of AI via its procurement of AI capabilities.

2.3. Sovereign capability
It is vital for Australia's economic and strategic security to invest in a sovereign AI capability. Indeed, Australia could distinguish itself internationally as the place where AI is used to give people a "fair go". We own this brand already. We emphasise the importance of maintaining the sovereignty of our data, as well as our home-grown AI innovations.

2.4. Transparency
AI will not be deployed responsibly without greater transparency into its deployment. Recently, tech companies like Twitter have increased fees to access their data greatly. This will prevent oversight from academia and NGOs. The government might therefore consider regulation to ensure affordable access to such for purposes of transparency.

2.5. Education
There is an important role for education and training of society to harness and leverage on AI. This can address both the public's fears and help to promote the responsible use of AI. Education needs to target everyone, from children (K to 12), to pensioners.

Detailed response

We also address in detail some of the specific questions in the public consultation.

Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?

3.0. The inclusion of the phrase “without explicit programming” in the definition of AI is problematic. It narrows the definition of AI to machine learning. It excluded much GOFAI (good old-fashioned AI). We recommend it be removed.

3.1. The definition that generative AI models “generate novel content” ignores what might prove to be the most impactful characteristic of large language models, viz., their ability to summarise and synthesise information.

3.2 We recognise that defining AI is hard, and that any working definition will evolve. Any regulation should take account of this.

3.3 The definition of machine learning as “the patterns derived from training data” is not accurate. Machine learning as a field of AI covers a broader view in terms of algorithms and systems.

What potential risks from AI are not covered by Australia’s existing regulatory approaches? Do you have suggestions for possible regulatory action to mitigate these risks?

4.0. A critical risk is that AI development may ignore human values, and impact human rights, privacy, consent, individual dignity, transparency, and inclusivity. Such impacts can often be addressed by keeping humans in (or -on) the loop. There are certain machineries and/or systems where full automation is widely acceptable and in fact, required to ensure a more fail-proof system. This includes a pacemaker, automated power production, and many more. In such cases, a safety switch is always needed when such systems are deployed, allowing a human agent to intervene.

4.1 One emerging area where ethical and legal frameworks will be especially important is in neurotechnology. In combination with brain-computer interfaces and novel sensing and wearable technologies, AI will make it possible to read people’s thoughts and even influence their actions, creating new challenges around privacy and autonomy.

4.2. We should recognise that AI could enable existing problems to grow in scale (such as misinformation), as well as present us with new, novel problems (such as deepfakes). However, a recent Australian effort also highlighted that human-AI cooperation, leading to hybrid intelligence, can also tackle misinformation and polarisation (Spina et al., 2023, CACM).

4.3. A critical risk emerges from AI policy as its implementation goes from theory to practice. The core assumptions about AI policy, which include 1) that AI is intelligent; 2) that more data is a requisite for better; 3) that AI itself can be ethical, can result in real harm if not challenged. Assumptions need to be recalibrated and adapted to the context of the policy application.

4.4. A black-box approach to decision-making using AI technology creates real risks of biases, discrimination, and unfairness to vulnerable populations. AI technologies without adaptations and tuning, are vulnerable to false interpretations of patterns of correlation as causal relationships.

Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or impacts.

5.0 Education must be at the forefront of Australia’s approach to responsible AI in order that we continue to invest wisely in our sovereign capabilities in this space.

5.1 Just like water safety and swimming skills, we must up-skill our young Australians across K-12 levels. Such initiatives must reach across all generations. A wide-ranging program of public information is required. Critical thinking has never been required with more urgency than it is now.

5.2 The government should explore formalising codes of good practice (see e.g., (Reid et al (2023)), as well as keep up to date with international standards.

5.3. Education and public awareness are needed to disseminate the limitations of AI. The risks come from the users of the AI technology being unaware of the assumptions behind it, and its effect on its applications in their context.

5.4. Government policy making should join forces with academics and tap into the recent development in causal AI.

Do you have suggestions on coordination of AI governance across government? Please outline the goals that any coordination mechanisms could achieve and how they could influence the development and uptake of AI in Australia.

6.1 Innovation can be encouraged via bridging mechanisms between government, industry, and academia, subsidising knowledge exchange. The outputs of these innovations will be of considerable value to civic culture at large.

6.2. There is an AI ecosystem within and between countries. Australia cannot operate separately from other jurisdictions. Equally, different stakeholders within Australia must together adapt existing regulations, and build new ones to avoid redundancies, conflicts, and loopholes in regulations, as well as create efficiencies from the relationships between interested parties.

Are there any governance measures being taken or considered by other countries (including any not discussed in this paper) that are relevant, adaptable and desirable for Australia?

7.0 The EU has drafted its "AI Act" which is referred to as a step closer to the first rules on Artificial Intelligence. It includes bans on biometric surveillance, emotion recognition, predictive policing AI systems, and the right to make complaints about AI systems. It remains to be seen which aspects of the proposals will be considered in Australia, but a discussion is welcome.

7.1 To get the right perspective, it is important to get a consensus on what values differentiate the Australian context.

7.2 One possible concern with these efforts is "over-regulation". It is important to look for research and capability opportunities rather than having a first impulse to regulate every issue.

Should different approaches apply to public and private sector use of AI technologies? If so, how should the approaches differ?

8.0 Public sector uses of AI should be, and should be demonstrably, for the Public Good. This is as opposed to being for the financial profit of shareholders. Like any other Public Good, this could operate at a net financial loss, subsidised by the taxpayer, like public hospitals for example.

8.1 Both public and private sectors should be targeted. Furthermore, related technologies and industries such as robotics, autonomous cars, the internet of things, and critical applications such as defense, education, and health should be framed under a general framework of ethical and moral uses of AI.

8.2 Regulations usually differ by line of business (public/compulsory/government versus private/voluntary) and jurisdictions. For example, see the discrimination regulations for insurance products (Frees and Huang 2023, Xi and Huang 2023) based on different social and/or economic principles. Generally speaking, there are more/stronger regulations for products/services provided by public sectors than private sectors. The same principles could apply to the AI (and non-AI) environment.

8.3 One option is adopting a principle/outcome/value-based approach (regulatory framework) rather than purely technique-based approach. For example, if the principle/outcome is fairness related, the difference between public and private sectors should be due to the difference in fairness notions - the public sector focuses more on the social aspects, while the private sector focuses more on the economic aspect. The difference in fairness notions leads naturally to differences in regulations that can affect the technologies (both AI and non-AI) involved in decision making. Note that this difference is not caused by AI technology itself, but by its underlying principles.

How can the Australian Government further support responsible AI practices in its own agencies?

9.0 Adopt a code of ethics similar to the IEEE Code of Ethics (see e.g., (IEEE P7000 standards)) or the Association for Computing Machinery Code of Ethics for the developers of AI across Government agencies.

In what circumstances are generic solutions to the risks of AI most valuable? And in what circumstances are technology-specific solutions better? Please provide some examples.

10.0 The risks cross over from AI being leveraged to compromise cyber security, to misinformation causing epidemics of false beliefs that might be hard to counter. In most cases, it would be helpful to lean towards human-centric, trust-based initiatives.

10.1 Robot interaction, and autonomous vehicles are two of the end-user critical applications that should be targeted. Also, the use of AI in social media generating behavioral addictions has been overseen and not discussed adequately. Another example is the use of AI in technology that directly impacts human lives (as in the case of online recruitment, in order to automate resume shortlisting), a context that requires technology-specific solutions to the risks posed.

10.2 The risks generated by the predatory use of AI technologies can be mitigated by government regulations. There are also risks caused by the naïve or ignorant implementation of AI to applications that affect humans directly. As an example, the black-box approach to AI in applications across health, HR, and governmental economic interventions has far reaching implications.

Given the importance of transparency across the AI lifecycle, please share your thoughts on: where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?

11.0 Transparency is critical, and companies are not sharing enough for proprietary reasons. Certainly, it is not transparent enough for the public. Transparency in disclosing the nature of the training datasets is important. Similarly, transparency in the evaluation data used to make claims about fairness and responsible behavior is also required.

Given the importance of transparency across the AI lifecycle, please share your thoughts on: mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

12.0 Transparency (linked to interpretability of AI/ML) is important, and the level of importance varies by application area. For example, for high-stake decision-making (say critical medical related decision making), the highest level of transparency is needed; while for translation tasks, a low-level of transparency requirement is needed. It will be helpful to clearly specify the specific level of transparency required (again principle/outcome-based) for different areas, this will automatically constrain the applicable model/techniques space to be used. In this case, for some high-stake decision-making areas, purely interpretable models are required, and black-box models are ruled out.

12.1 It is also important to enhance the AI literacy of all stakeholders, as some black-box models claim to have explanation tools based on some literature, which, however, are vulnerable to adversarial attacks and based on strong assumptions. It can be dangerous to apply these explanatory tools for high-stake decision making.

12.2 Another important aspect for transparencies are Machine Learning operations (MLOps), ensuring full documentation of the training data, features, experiment setup, etc. and enabling algorithmic auditing if things go wrong, and enabling the traceability of decision-making processes (see, e.g., Raji et al. 2020 and Sokol & Flach 2020). Transparency in disclosing the nature of the training datasets and the model training is vital.

12.3 We acknowledge that while there are some comprehensive efforts (see, e.g., Sokol and Flach, 2020) towards explainable AI (XAI), the tools are currently incomplete.

Do you have suggestions for: whether any high-risk AI applications or technologies should be banned completely?

13.0 This is a discussion that we need to start to have in an ongoing fashion. Some AI technology is specially concerning, such as facial recognition that identifies sexual and political preferences. Whether accurate or not, the potential for catastrophic misuse is obvious. A counterpoint is that instead of a ban, regulation is a more pragmatic approach. For example, one concerning area is social media used by children and teenagers.

13.1 One consideration is to use a principle/outcome-based approach rather than technique-based approach for decision making. For each application area, one could specify clearly the principles (based on the legal framework) required. If an AI-based application were high-risk, it would violate one or more principles for this specific area of application. Instead of banning high-risk AI applications, we should spend greater effort on making our legal framework (with non-legal guidance resources) more comprehensive (with no or fewer gaps), so that high-risk AI-based applications are ruled out naturally. Moreover, the outcome-based approach provides a consistent framework for both AI and non-AI techniques.

Do you have suggestions for: criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

14.0 Criteria could include when AI is used by vulnerable populations such as children, seniors, and people with disabilities. Any application involving potential persuasion and manipulation should definitely be regulated (gambling, education, social connection, and online shopping for example).

14.1 Industries that are well-regulated have in-built mechanisms to safeguard against risky technology. When identifying areas where there is a need to exercise more care, it will be useful to consider industries that are not as well-regulated.

What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

15.0 Enhanced AI (including responsible AI), literacy at all levels of society (from the general public to professionals), technicians, and managers (all stakeholders). Many people don't understand (responsible) AI and hence fear to apply it, because it is a mystery to them. One suggestion then is to implement a hands-on AI education program for students and professionals. Another suggestion is to invest more effort in the public understanding of (responsible) AI.

How would banning high-risk activities (like social scoring or facial recognition technology in certain circumstances) impact Australia's tech sector and our trade and exports with other countries?

16.0 A well-regulated tech sector that bans certain abhorrent and high-risk activities could be an advantage and a selling point.

Do you support a risk-based approach for addressing potential AI risks? If not, is there a better approach?

17.0 Risk-based approaches look well designed, especially if they consider both long- and short-term risks. This approach is currently used in several domains, e.g., health. There can be more concrete and comprehensive frameworks of evaluation. For example, when the guideline says, "ensure robustness, accuracy and cybersecurity", it would be more useful to specify how these requirements can be measured and enforced, and what would be the recommended process to establish these evaluation frameworks.

17.1 We would like to caution that poorly targeted and carelessly framed regulation can have unintended consequences, so it must be designed with care.

17.2 A Risk Based approach can provide a suitable framework to avoid the devastating policy mistakes that come from the use of AI to predict human behavior. For example, risk-based approaches can adapt AI algorithms to areas of health and development programs and build causal models to better explain why people behave the way they do.

17.3. What constitutes risks can get outdated easily given the fast pace of AI technology development. It is essential to consider how this approach can be updated over time easily.

17.4 It is also critical to ensure that any risk-based approach does not discourage basic AI research and does not discourage home-grown AI innovations from thriving. A responsible AI Framework should be one of the safeguards in place, in addition to compliance to Human Research Ethics frameworks.

What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?

18.0 One limitation is that people disagree on what constitutes harm, and it is potential harm that constitutes the risks(s) in question. For example, most of the dialogue surrounding AI in criminal sentencing decisions recognises only the potential harm caused by predictive error. However, there is also the potential harm arising from the lack of accountability or provision of reasons presented by any suitably complex and opaque predictive AI, irrespective of any predictive accuracy. The danger here is that we might harm the tapestry of our social fabric.

Is a risk-based approach better suited to some sectors, AI applications or organisations than others based on organisation size, AI maturity and resources?

19.0 Certain sectors require this approach. For instance, in terms of robotics, it is promising to see generative AI and Reinforcement learning applied in training sophisticated robots. However, this should be done in controlled environments with exhaustive testing before having commercial products.

What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?

20.0 Before classifying an issue such as data privacy as low or high risk, it needs to be contextualised with the corresponding use-case such as in medicine, or for use by children. One can only deal with the issues within each use-case separately.

How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?

21.0 Risks are faced not only by end users of AI, but also by those who work to maintain the AI infrastructure itself. We must consider also the risks faced by data curators as well as all of those whose labour is extracted to maintain AI systems. Furthermore, rare earth minerals and the huge amount of water needed for machine-cooling, are not without costs. As with any resource extraction, the risks of such extractions must be discussed and factored transparently.

21.1 LLM's are not infallible. There is a risk that the reality of this fallibility could be lost in the enthusiasm for AI itself, resulting in a considerable amount of entrenched misinformation. Mitigating this will propagate honest discussion, the production of innovative research, and better education.

21.2. A comprehensive risk-based approach is one of the many useful tools to examine general AI systems.

21.3 Also consider using a veracity-based approach where applicable (see, e.g., (Luczak-Rösch et al 2023)). To ensure and prove veracity, people, organisations, and applications preregister secure identifiers related to relevant claims while maintaining control over the data related to the claim. In this setting, the diverse cultural and social contexts can also be incorporated appropriately.

Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to: public or private organisations or both; developers or deployers or both?

22.0 A risk-based approach to responsible AI should comprise both voluntary and regulated components. High-risk activities should be governed by regulation, whilst lower risk activities could be managed through systems of voluntary self-regulation. The identification of activities as being of high, medium, or low risk, should be made with regard to transparent criteria, by experts whose appointment is motivated by accountability. This approach should reach across both public and private organisations, and across both AI developers and AI deployers.

22.1 It is important that these regulatory standards are consistent with those such as the guidelines of professional associations - ACM, IEEE, AAAI - among others. Some representative guidelines include "*Contribute to society and to human well-being, acknowledging that all people are stakeholders in computing*" and "*Be honest and trustworthy*" (ACM, 2023).

22.2 The regulatory framework must be agile or plastic enough to keep pace with the high speed of AI development.

In conclusion

Please do not hesitate to contact Dr Haris Aziz at ai.director@unsw.edu.au should you wish to discuss any issue raised in this submission.

References:

ACM. (2023). ACM Code of Ethics and Professional Conduct. ACM Code of Ethics and Professional Conduct. https://www.acm.org/code-of-ethics

Alistair Reid, Simon O'Callaghan, and Yaya Lu. 2023. Implementing Australia's AI Ethics Principles: A selection of Responsible AI practices and resources. Gradient Institute and CSIRO.

Australian Alliance for Artificial Intelligence in Healthcare (AAAiH), A Roadmap for AI in Healthcare for Australia, Report, 2021

Australian Human Rights Commission, Human rights and technology: final report, September 2021 https://humanrights.gov.au/our-work/rights-and-freedoms/publications/human-rights-and-technology-final-report-2021

EU Commission (2020) White Paper on Artificial Intelligence—A European Approach to Excellence and Trust, COM (2020) 65 final, Brussels: European Commission.

Frees, E.W. & Huang, F. (2023). The Discriminating (Pricing) Actuary, North American Actuarial Journal, 27:1, 2-24, DOI: 10.1080/10920277.2021.1951296
Hogenhout, L., A Framework for Ethical AI at the United Nations, United Nations, 2021.

IEEE P7000 standards? https://standards.ieee.org/industry-connections/ec/autonomous-systems

Markus Luczak-Rösch, Matthias Galster, Kevin Shedlock: The Veracity Grand Challenge in Computing: A Perspective from Aotearoa New Zealand. Commun. ACM 66(7): 67-69 (2023)

Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, Parker Barnes: Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. FAT* 2020: 33-44

Kacper Sokol, Peter A. Flach: Explainability fact sheets: a framework for systematic assessment of explainable approaches. FAT* 2020: 56-67

Spina et al., 2023, CACM, https://cacm.acm.org/magazines/2023/7/274056-human-ai-cooperation-to-tackle-misinformation-and-polarization/fulltext

Walsh, T., Levy, N., Bell, G., Elliott, A., Maclaurin, J., Mareels, I., & Wood, F. (2019). The effective and ethical development of artificial intelligence: An opportunity to improve our wellbeing. Australian Council of Learned Academies.

Xi Xin & Fei Huang (2023) Anti-discrimination Insurance Pricing: Regulations, Fairness Criteria, and Models, North American Actuarial Journal, DOI: 10.1080/10920277.2023.2190528