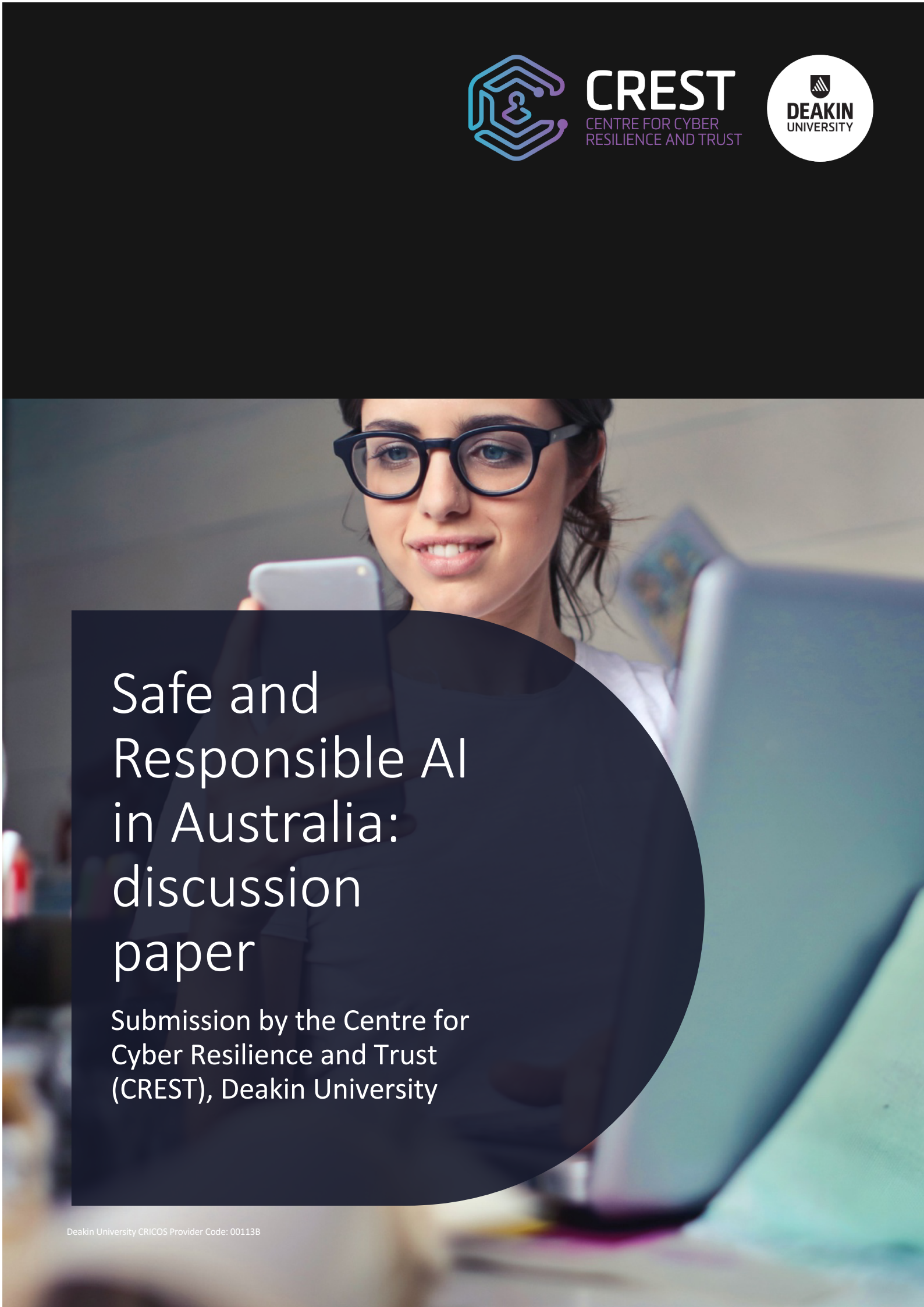

Safe and Responsible AI in Australia: discussion paper

Submission by the Centre for Cyber Resilience and Trust (CREST), Deakin University

Safe and responsible AI in Australia: discussion paper

Submission by the Centre for Resilience and Trust (CREST), Deakin University

The Deakin University Centre for Cyber Resilience and Trust (CREST) is pleased to make this submission in response to the safe and responsible AI in Australia discussion paper. Our submission is underpinned by the multi-disciplinary research expertise of CREST with contributions from scholars across technology, social science, law, and policy.

About Us

CREST brings a multi-disciplinary focus to the changing landscape of cyber harms and the extent to which people, organisations, and communities are dependent on the growing digital economy. The term 'cyber resilience' encompasses but extends beyond the notion of conventional understandings of 'cyber security'. Our focus on cyber resilience exists at individual, organisational, and societal levels, and emphasises a need to move beyond a singular focus on preventing cyber security incidents to also anticipating, protecting, detecting, mitigating, disrupting, and recovering from them. CREST seeks to analyse the role of trust in the design of systems, cyber security technologies, and the capabilities of users. We examine mechanisms of trust in both technology and humans – how these are created and what they can achieve – in advancing cyber resilience.

CREST aims to utilise multi-disciplinary expertise to design state-of-the-art cybersecurity solutions by responsibly leveraging emerging technologies such as artificial intelligence, blockchain and quantum computing whilst also enabling a comprehensive understanding of the role played by the human factor and governance in cyber security. We employ the term 'human factor' to consider a broad suite of attributes relevant to cyber resilience at the individual, institutional, and societal levels. This conceptualisation includes individual human behaviours, as well as the social structures that enable collective action by groups and communities of various sizes, and the diverse public and private interventions that shape societal responses. We also seek to directly extend our focus to the diverse actors responsible for cyber harms, and the institutions and regulatory approaches necessary to prevent, minimise, and recover from such harms. Our focus on the human factor also extends to the notion of 'usable security' – ensuring that cyber security technologies are designed to be user-centric, inclusive, and affordable.

CREST adopts multi-disciplinary approach to these critically important knowledge gaps through five interrelated areas of impact:

- Advancing cyber security technologies
- Securing data and infrastructure
- Promoting cybersafe behaviours
- Disrupting cyber harms
- Harmonising cyber governance

Contributed by:

Prof Gang Li, A/Prof Lennon Chang, Prof Shiri Krebs, Dr Nayyar Zaidi, Prof Chad Whelan, and Prof Robin Doss.

Safe and responsible AI in Australia: discussion paper

Definitions

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

Response:

Definition of AI

The definition of artificial intelligence (AI) needs to be broad, as it encompasses a wide range of technologies and applications. AI systems can be used to perform a variety of tasks, including predictive modelling, pattern recognition, and decision making. Some AI systems are designed to be predictive, while others are not. For example, some AI systems are used to identify patterns in data without making predictions about the future. These types of AI systems are often referred to as "nowcasting" systems.

In this sense, the definition of AI in this document just focuses on engineered AI systems, which are designed and developed by humans. This contrasts with Artificial General Intelligence (AGI), which is a hypothetical type of AI that would have the ability to learn and reason like a human being. AGI is not yet possible, but it is a long-term goal of many AI researchers.

ML Definition

The definition of machine learning is deficient, and the use of “machine learning algorithms” while attempting to define the term (i.e., machine learning) is problematic.

A textbook definition of the term can be found in Machine Learning (Tom Mitchell, 1997) and is as below:

“A computer program is said to **learn** from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Any agreed and accurate definition needs to emphasise learning from experience and not on “patterns derived”, “prediction” or “decision-making”.

Potential gaps in approaches

2. 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?

Response:

Based on recent work by Dr Jayson Lamchek and Professor Shiri Krebs (one of the co-authors of this submission),[1] gaps responsible AI framework, in particular – risks that are not currently addressed by existing frameworks – can be categorised into four pillars: ethics, society, law and security. While some concrete areas under each pillar may be covered by existing data privacy laws and data governance guidance, existing laws, regulations, and guidelines are too specific and not inclusive enough to cover existing gaps.

In particular, under the ethics pillar, there is currently insufficient protection against the harms caused by AI to the **autonomy** of both decision-makers and users of these technologies (for example, as a result of embedding AI-based data filtering algorithms or automating parts of the decision-making cycle.) Other important ethical gaps relate to the lack of clear guidelines regarding the **explainability and understandability** of the AI elements in human-machine interactions, as well as to **accuracy** gaps resulting from the use of synthetic or partial data sets for training AI models and their potential incompatibility with real-world data and the lack of transparency in relation to the type of training data used.

Under the law pillar, existing laws fail to address issues relating to **compliance and legal responsibility** for harms caused by AI systems. Human rights protection is also partial, with privacy as the main protected right in this context. However, additional consideration of other human rights is still lacking – for instance, dignity and equality.

Under the social pillar, existing frameworks do not provide solutions to problems such as **employment and de-skilling** (with loss of jobs as well as loss of human skill), as well as sustainability (with little to no concern given to the **environmental costs** of training and tuning machine learning and AI models.)

Finally, under the security pillar, existing frameworks are yet to develop a tool to evaluate the trade-offs of **security and functionality**, with additional security requirements often leading to loss in functionality, as well as loss of access to services, particularly with regards to vulnerable communities (including people with disabilities). While this trade-off is avoidable, secure by design principles and approaches are needed to ensure higher levels of assurance.

3. 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.

4. 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.

Responses (to questions 3 and 4 combined):

Based on the work by Dr Jayson Lamchek and Professor Shiri Krebs mentioned above, we propose an AI risk-assessment framework that is inclusive of the four pillars mentioned: ethics, society, law, and security.

[1] Jayson Lamchek and Shiri Krebs, 'Cybersecurity Research and Society: Considerations for Researchers and Human Research Ethics Committees' in Bruce Smyth, Michael Martin, and Many Dowling, eds., *Routledge Handbook of Human Research Ethics and Integrity in the Australian Context* (forthcoming, Routledge, 2023).

Research projects developing technological solutions designed to improve the cybersecurity of critical infrastructures – such as electricity or communication systems – are often reviewed for their research merit and security features, without triggering a human research ethics review process that considers broader social and ethical implications. This is because their design does not trigger compliance-based standards or regulations.

However, such projects often do have significant social effects, which tend to remain unexplored and unaccounted for. Many cybersecurity research projects, especially those designed to protect critical infrastructures, involve making many complex and often invisible value-judgments, as well as balancing between competing social interests and needs. The resulting cybersecurity tools have distinct impact on our lives, often affecting individuals' human rights, social values, and the environment. As these effects go beyond the current legislation or existing human research ethics review standards and processes, research institutions must adapt their AI governance and project design processes to reflect the potential impact on society.

As identified by Lamchek and Krebs, we recommend a review process for any new utilisation of AI technologies that includes all four pillars mentioned above: ethics, society, law, and security. Krebs and Lamchek further include specific values and considerations under each of these pillars to guide the review process.

Implementing this recommendation further suggests the need for (a) expanding human ethics committees membership and expertise to provide a deeper review of AI applications and their potential effects on humans and societies; (b) enhancing the review process to include broader social, ethical and legal considerations, including sustainability, explainability and respect for individuals' dignity and autonomy; and (c) implementing effective ways to cross disciplinary divides within research teams and between research teams and their stakeholders, users, and society members more broadly.

Finally, we emphasize the need to go beyond abstract principles and to include actionable items for both developers and reviewers, designed to change research culture and environment. This cultural change – focused on a holistic review of AI applications within the environment they intend to operate in – further necessitates self-reflection and a continuous deliberative process that is inherently linked to society members and broader social values.

Responses suitable for Australia

5. 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?

Response:

The consultation paper provides a good outline of measures adopted by various countries, as well as measures adopted by the Australian government. As AI is an application that can be used in everyday life, it is not solely a matter for the government to govern. Public and private collaboration is needed to govern this new space. As outlined in the consultation documents, we see a wide range of ways to govern AI, from voluntary self-regulation to strict laws. Similar to the governance of cyberspace, where debate exists whether voluntary or compulsory regulation would be more effective[2], it is hard to say with certainty which is the most effective way to regulate AI.

[2] Chang, LYC. (2012). Cybercrime in the Greater China Region: Regulatory Responses and Crime Prevention across the Taiwan Strait. Cheltenham: Edward Elgar.

Safe and responsible AI in Australia: discussion paper

However, learning from our experience with cyber security, we suggest that the Australian government develop a national strategy on AI as a high-level guideline for the development and use of AI in Australia and proactively participate in the negotiation of AI standards and guidelines, regionally and globally. This is especially important when it comes to national security and safety as we see the potential harm that AI can bring to national security and democracy.

It is also suggested that Australia takes part in or leads a regional and/or global alliance on the use and the impact of AI, so as to provide a forum for the regular review of AI's development and the approaches that need to be adopted to tackle issues identified.

AI related policies from governments:

- USA: nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf
- UK: AI in the UK: ready, willing and able? - government response to the select committee report - GOV.UK (www.gov.uk)
- EU: Coordinated Plan on Artificial Intelligence | Shaping Europe's digital future (europa.eu)
- Singapore: AI.SG: New National Programme to Catalyse, Synergise and Boost Singapore's AI Capabilities (smartnation.gov.sg)
- Japan: ＡＩ戦略2022の概要 (cao.go.jp)
- India: National Strategy For Artificial Intelligence (indiaai.gov.in)

Target Areas

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

Response:

While the public and private sector might use AI technologies in different ways, it is crucial that all use of AI follows rules that enhance trust and confidence not just in AI but in delivery of services by both government and business. This should include the ethics principles in Australia's AI Ethics Framework, such as transparency, contestability, accountability, reliability, safety, fairness, privacy and data protection. It is also critical to ensure that the use of AI does not threaten democracy, nor mislead society.

However, the public sector has a greater responsibility to lead and to ensure that AI does not have a negative impact on society resulting from a biased database or the algorithms used to generate information. A more stringent approach should be applied to the use of AI by the public sector to reduce the risk of neglecting marginalised groups and communities, such as the first nations and migrant communities. There is also a need to have a strongly regulated approach to using AI by national security related government entities. Clear guidelines need to be set to prevent potential risks that might come with the use of AI and to protect the national interest. If necessary, the Australian government should ban the use of AI in certain government entities and/or ban the generative AI applications if there is evidence of biased information in order to promote certain ideology or propaganda.

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

Response:

Currently the guidelines and regulations are built around the use of data, however, concrete directions and recommendations can be made for various models specific to each domain. E.g., a deep neural network which is not interpretable is okay to be deployed at Airport Security gate to scan an incoming passenger at the gate but should not be used for determining a tax fraud or determine credit worthiness.

Each model has its strength and weakness; therefore, each model should be recommended for different scenarios. A process or policy should be formulated for cases when a least desirable model is still required. As an example, for a tax fraud scenario, it can be the case that explainable models are just too complex to train given they might require loading entire data in the memory, while a non-explainable model such as deep artificial neural network is desirable due to its handling of large-scale complex data. Guideline on the use of various AI models in government agencies will go a long way in building effective solutions in a responsible fashion.

8. 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.

Response:

This is related to our response to question 7. A generic solution to mitigate AI risks is valuable when it comes to devising risk mitigation strategies pertaining to public and masses. For instance, risk associated to a technology that can create deep fake news or images does warrant a generic solution. However, for specific sectors, a technology-specific solution may be required e.g., risks associated with falsely predicting someone's credit score, or risks associated with showing an advertisement to a web browser. The risks associated with various sectors can be identified in consultation with domain experts.

9. Given the importance of transparency across the AI lifecycle, please share your thoughts on:

    a. where and when transparency will be most critical and valuable to mitigate potential AI risks and to improve public trust and confidence in AI?

    b. mandating transparency requirements across the private and public sectors, including how these requirements could be implemented.

Response:

Transparency is a core requirement from any AI-based application. The technology, including the way it operates and arrives at conclusions or generate outputs must be transparent to human beings who use, interact with or are affected by the technology. The requirement of transparency can be better understood through two related AI ethics principles: understandability and explainability.

A number of AI ethics guidelines promote the importance of explainable AI, which Ryan and Stahl (2021) termed understandability.[3] This principle requires that organisations should understand how their AI-based systems and technologies work and explain the technical functioning and decisions reached by those technologies, whenever possible (Floridi et al., 2018).[4] The data, algorithms and the decisions that will be arrived at by those processes, and the actions taken by AI should be comprehensible by human beings (European Parliament, 2017).[5] A related principle – explainability – requires that organisations document how their AI-based technologies or automated processes reach certain decisions and be able to reproduce them for audits.

Both these principles mean that AI-based processes must be transparent to the humans involved and affected by the technology. Indeed, existing AI ethics guidelines require developers to maximise transparency regarding access to and use of personal data (United Nations Development Group (UNDG), 2017), as well as informing users that AI is being used (Ryan and Stahl 2021). Transparency of AI products and processes encompass not only the design of the technology and its functionality, but also the potential harmful effects it may generate. Transparency is important because it allows individual users to make informed choices about sharing their data and using AI.

Therefore, we recommend that any AI ethics framework will require embedding transparency amplifiers into the design of new AI applications, assuring that users of the technology understand how it functions, as well as explaining the limitations of the technology-generated outputs and advice. Such transparency will support not only correct utilisation of the technology but will also enhance user trust in the technology and cybersecurity workers' skill.

[3] Ryan, M., & Stahl, B. C. (2021). Artificial intelligence ethics guidelines for developers and users: Clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, *19*(1), 61–86.

[4] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[5] European Commission. 2021. "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts." https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

10. Do you have suggestions for:

    a. Whether any high-risk AI applications or technologies should be banned completely?

    b. Criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

Response:

We believe AI technology is as good as the data it is trained on. Therefore, regulations should be set for companies to do due diligence in checking the authenticity of the data. They should specify the source of the data as well as its salient features. Companies should be obligated to have a rigorous testing schemes set up before releasing their model (even beta phase models). Rather than banning the technology, use of beta-models should be either restricted or invigilated by an authorized authority or banned all together -- as most of these models are aimed at gathering more data to perfect the model. Companies should be banned from using consumer data, without their explicit consent.

Banning a technology is never a solution, however, proper training in the form of employee training as well as public awareness in form of consolidated programs is a much better strategy. Frameworks should be built around detection of misinformation and dis-information.

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

Response:

There is still insufficient training and awareness raising activities relating to the use of AI. We even see some governments banning the use of AI. These won't help increase public trust in AI deployment. While AI is becoming part of our everyday life, it is crucial to provide more information about AI through awareness raising campaigns and education. To design these programs, it is important not only to teach people how to use AI, but to let people be aware of the advantages and disadvantages that AI might bring to society. Only when users are familiar with AI and the issues related to AI, they will start to build trust in this new technology.

Implications and infrastructure

12. 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?

Response:

Given that "high-risk activities" stem from the application of the technology rather than the technology itself, it is unclear how such an approach could be beneficial or even practical. However, such an approach to ban high-risk activities could stem innovation which could have flow on effects on Australia's tech sector, trade, and exports. For example, it might hinder innovation and development of these technologies and prevent Australia from playing a leadership role both in the Indo-Pacific and globally for responsible use of the technology and protection of a rule-based order. It might also result in fracture of relationships with other countries that may be using the technology for what Australia might view "high-risk" activities. However, it is important that Australia takes into consideration national security, civil liberties and the protection of democracy while considering the use and regulation of AI.

13. What changes (if any) to Australian conformity infrastructure might be required to support assurance processes to mitigate against potential AI risks?

Response:

A range of infrastructure needs to be implemented or revised to better support AI safety assurance processes. This includes:

Laws: Updates to liability laws and cybersecurity laws may be needed to determine accountability when AI systems cause harm. However, care must be taken not to over-legislate, given the fast pace of progress in the field of AI.

Standards bodies: National standard organizations, such as Standards Australia, can create standards focused on AI safety, testing, and transparency. This will help promote best practices and ensure that AI systems are developed in a safe and responsible manner.

Safety regulations: Existing safety regimes, such as those administered by the Australian Safety and Compensation Commission, the Australian Competition and Consumer Commission, the Department of Health, and the Department of Transport and Infrastructure, may need to be updated to include requirements for the review, testing, and approval of AI systems. This is especially important for high-risk applications, such as those used in medicine or transportation.

Auditing processes: Third-party auditing and testing protocols can be developed to evaluate the properties of AI systems, such as their reliability, security, and fairness. Having standardized auditing frameworks will help to provide assurance that AI systems are safe and effective.

Risk-based approaches

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

Response:

Risk-based approaches have been used successfully in banking sector (money laundering, etc). A risk-based approach to AI risks does makes a lot of sense. It is already adopted in Canada and should be adopted in Australia. We do not see a better way of managing risk given the different nature of various AI algorithms and their formulations.

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

Response:

The main benefit of risk-based approach is that it allows a sector specific as well as technology specific risk analysis. E.g., use of deep artificial neural network models can be low when doing computer vision tasks while can be high when used in sectors where more interpretability is required. The risk levels can be modified based on organization's maturity as well as their track record e.g., what internal processes have they set-up during model training and testing? What models are being used? What is the quality of data they have obtained? How do they test the model for covariate and concept drifts, as well as controlling for fairness and privacy? Once processes are formulated, one can easily identify the level of risk associated with AI model.

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

Response:

See our response to Q15. Yes, risk-based approach better suits some sectors, technologies as well as organizations, as the level of risk depends greatly on the quality control process of the organization building AI solutions as well as the technology they are

using. A Risk-based approach is advantageous as it can consider different elements that are involved during the building of AI solutions.

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

Response:

The proposed framework is a good starting point for the development of AI safety rules. However, there are some additional elements that could be considered, such as

- Validation protocols for intended behaviour, especially for autonomous systems. This would help ensure that AI systems are designed and developed to behave in a safe and predictable manner.
- Registration or certification requirements for providers of high-risk AI services. This would help ensure that these providers have the necessary expertise and experience to develop and deploy AI systems safely.

The proposed framework could also be enhanced to include,

- Notices: The timing and scope of notices should be considered. Notices may be most useful if they are provided prior to the use of AI.
- Explanations: Explanations are critical, but guidance could be clearer on what constitutes a satisfactory explanation of an AI system's decisions. Explanation methods vary widely, so it is important to have a clear definition of what is expected.
- Monitoring: Ongoing monitoring is good, but more guidance is needed on specific techniques such as algorithm auditing, bias testing and performance metrics. Confidence that the AI is working as intended needs to be measured and monitored on an ongoing basis. If necessary, the AI needs to be retrained.

18. How can an AI risk-based approach be incorporated into existing assessment frameworks (like privacy) or risk management processes to streamline and reduce potential duplication?

Response:

NA

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

Response:

A risk-based approach to general purpose AI systems would involve identifying and mitigating the risks associated with these systems, in particular:

- Misinformation and disinformation: LLMs and MFMs can be used to generate racist messages or biased statements, or to spread misinformation and disinformation, which can have a negative impact on society.
- Breaches of privacy or confidentiality: There are potential pitfalls in training these models, such as IT security issues, intellectual integrity and property protection concerns, and potential breaches of privacy or confidentiality.

The mitigation of them needs to consider elements, such as:

- Risk assessment: Identify potential risks in areas such as bias, creation of harmful content, spread of misinformation, breaches of privacy, hacking, misuse of intellectual property, etc. Risk assessments would evaluate factors such as the amount of data used, model size, training approaches, fine-tuning capabilities, etc. that determine the potential damage they could cause.

- Safety by design - Address risks through safety by design techniques such as controlled training data, user controls, technical safeguards against misuse, advisory restrictions on use cases.
- Iterative adaptation - Continuously re-evaluate risks or robustness of the system as capabilities evolve and new use cases emerge.

20. 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:
    a. public or private organisations or both?
    b. developers or deployers or both?

Response:

NA