RAND AUSTRALIA

THE RAND CORPORATION (AUSTRALIA) PTY LTD
ABN 60 600 246 026
PHYSICAL LOCATION: UNIT G8, 65 CANBERRA AVENUE
GRIFFITH, ACT 2603 AUSTRALIA
POSTAL ADDRESS: SUITE 24, 2 ENDEAVOUR HOUSE
CAPTAIN COOK CRESCENT
MANUKA, ACT 2603 AUSTRALIA
PH: 02 6232 6972
EMAIL: RAND_AUSTRALIA@RAND.ORG

Department of Industry, Science and Resources
AI Discussion Paper Consultation Hub

To whom it may concern,

The RAND Corporation, including RAND Australia, is undertaking a range of studies associated with risks inherent with the development of artificial intelligence. We have shared some of this work with the Department, and intend to continue this provision of information as desired. In relation to the request for feedback on the Sup*porting responsible AI: discussion paper*, attached are some initial thoughts on select questions, although we expect to provide more input as our studies progress further.

My lead on artificial intelligence in the Australian office is Dr Austin Wyatt, awyatt@rand.org.

Kind regards,

Andrew Dowse
Director, RAND Australia
Office: +61 2 6232 6972
Email: adowse@rand.org

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

The definition of Artificial Intelligence utilised in this discussion paper would be better classified as task-based or narrow AI. Whilst limiting the definition in this manner is arguably useful for the purposes of the discussion paper, it is still worth noting the limitation to task-based AI.

The other element of the definition that is worth questioning is that the limitation to systems that generate "predictive outputs" for a "given set of human-defined objectives". Again, while this limitation has some value for centring the discussion on near-term systems, it is worth noting the risk that it places future technologies outside the purview of the discussion.

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?

One measure not mentioned explicitly in the document is data localization, geographic restrictions on the storage of certain data types. GDPR has some data localization components that may be of interest to Australia. There are trade-offs associated with these data localization policies. One the one hand, data localization laws can protect the privacy and security of Australians and limit the efficacy of specific types of information attacks (e.g., personal information stored abroad can be used to micro-target individuals or populations with specific messages designed to with tailored messages that may be harmful). On the other hand, if the data used in AI models exclude information about Australians, their outputs may not be as relevant and could potentially be harmful. This could be particularly relevant to populations, such as indigenous Australians, that are not as widely represented abroad. AI models developed on foreign population would likely grossly underrepresent some Australian populations and the outputs of such models could exhibit significant bias in ways that are harmful when applied on the population of Australia. For example, an AI model used for identifying a specific health condition from biomarkers may not be effective for indigenous Australians because their records were not included in the training data. Such a model might provide inaccurate results or exhibit bias. However, data localization laws do not necessarily mean forgoing the benefits of AI models developed abroad. In some cases, models trained on foreign populations could be fine-tuned using Australian data that were compliant with a data localization policy. This would require cooperation with the foreign AI developers and resources for the development of Australia-specific results.

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

Yes, different approaches to regulating the use of AI should be taken by the public and private sector. While the same standards should be imposed on public and private sector actors for ensuring the security and protection of sensitive data (including source code, training data, and AI models) the mechanisms for ensuring those standards are maintained should reflect the differences between public and private sector actors.

As with cyber security, security regulations for private sector use of AI should balance enabling innovation with protecting against intrusion or malicious use. Industry actors

should be brought along with government in the development of regulation. Ethical codes of practice for private sector developers and users would also be a valuable step.

Public users, including government, law enforcement and military actors, should be required to undertake more stringent measures. In addition to more stringent access monitoring and data security provisions, regulation should limit the capacity of AI to make unsupervised decisions. There also need to be mechanisms for training human staff members and holding them accountable for foreseeable harms.

Here the concept of Meaningful Human Control, originally developed in relation to Lethal Autonomous Weapon Systems, is highly informative. A human operator in a government agency must be trained and knowledgeable about the AI technologies they utilise in making decisions that impact members of the public. For example, biases in the data used to train AI technology to support bail decisions makes it more likely that minorities are discriminated against, unless the magistrate is aware of, and knows how to recognise, that risk.

Regulation should also differentiate between governing the use of AI-enabled technologies and their development, including the curation of training data sets. While this is made more complicated by the proliferation of non-Australian owned technology in this space, imposing stringent data hygiene and safety requirements on developers responsible for training machine learning-based AI would have a significant flow-on effect, limiting the potential harm of such systems regardless of whether they were eventually used by the public or private sector.

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

This is a question that would benefit from a more detailed review. However, valuable first steps that the Department could take immediately would focus on publicising the risks and benefits of AI and limiting the potential for its misuse (whether recklessly or maliciously) by members of the APS. For example, training in how to ethically utilise generative AI for public information campaigns would encourage the responsible use of such technologies while limiting the risk of misuse. The development of a secured internal version of a commercial LLM could allow the Department to familiarise senior leadership with the technology while providing a controlled sandbox for experimentation and training by the APS with lower risk.

Another mechanism would be for the Government to provide training for APS members that intend to utilise AI-enabled technologies, especially generative AI models and Automated Decision Making programs. Humans have a tendency to overly trust in technology once it diffuses and matures, this automation bias can lead trained and experienced humans to trust in the technology rather than their own judgement, amplifying the risk of unanticipated consequences, for example via malicious actor interference or training data bias. Mandating the provision of some level of familiarisation and ethics training for APS staff that work with these technologies would limit these risks.

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.

Generic solutions can be suitable for mitigating the risks of AI in certain cases. For example, significant regulatory value could be achieved by imposing generic ethical standards on developers of AI-enabled systems through incentivising the inclusion of ethics training by tertiary education providers. Other examples would include the imposition minimum standards for data security and training for employees working with AI technology, mandatory reporting of data breaches or significant malfunctions (perhaps using the Privacy Act as a model), or imposing legal liability on those who maliciously or recklessly leverage generative AI for harmful purposes (i.e., maligning ethnic groups, cyber-bullying, or intellectual property theft).

The main areas where specific regulations are more valuable are those where the use of AI presents a unique scenario, can create harm in the physical world, or places individuals at risk. For example, AI-enabled technologies that support medical decision-makers or insurance firms, AI for self-driving cars or automated industrial equipment, or the use of automated decision-making programs for tax or legal liability review. Each of these would be cases where a combination of a unique use case and a higher risk of human harm would require specific regulation. In such cases policy makers should seek the input of professionals within that sector as well as technical AI experts in order to ensure that the resultant regulation is fit for purpose.

Moreover, consideration should be given to situations in which unintended and second order consequences are difficult to contain. This would include applications in cyber and biological domains, as well as generative AI.

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?

Because bias in the data used to train AI systems can lead to AI outputs that reflect and even exacerbate those biases, it is important for users of AI for critical tasks to understand the data sources, the biases in those sources, and the extent to which those biases are reflected in the outputs.

Transparency requirements are also particularly important for uses cases where automated decision-making programs are leveraged to inform decisions that are politically sensitive or present a significant risk of harm. While the robo-debt scandal is a particularly obvious example, others would include corporate use for targeted advertising of alcohol, informing patrol patterns for law enforcement, the denial of bail, or the division of education funding.

Similarly, the use of generative AI and LLMs for mass advertising or influence operations (including political campaigns) present a grave risk of harm to the health of public discourse. The engagement of such technologies for government messaging, corporate advertising, and political campaigns should therefore impose additional transparency and reporting requirements on users.

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

One approach would be for a common set of data quality measures to be developed and applied to data sources and AI model outputs. These measures would need to include both

overall accuracy characteristics and also accuracy for sub-populations or groups. These measures should be developed in consultation with the users of the AI systems and the people who will be subject to the outputs of the systems.

10B: Do you have suggestions for criteria or requirements to identify AI applications or technologies that should be banned, and in which contexts?

This is fundamentally about what parameters are used to define what constitutes a high-risk AI application, or perhaps more accurately, if the intention is to identify applications that are to be banned, an unacceptable risk AI application. Existing conceptualisations rely largely on the impact or projected threat of an application including risks that are likely soon to materialise or have already materialised. There are arguably four categories of risk identified by threat or impact, namely:

(1) Unacceptable risk, which is that which contravenes fundamental rights.

(2) High risk, which may adversely affect human health and safety or fundamental rights.

(3) Limited risk, which imposes requirements for transparency in certain circumstances so users know it is a machine that they are interacting with.

(4) Minimal risk, which allows other types of applications to be legally developed.

Alternatively, they may be identified based on the intended purpose of the application. Remaining cognisant that this might change, typologies of AI applications are used. This approach however, will likely be ad hoc and reactive. Moreover, the enumeration of high-risk AI systems does not mean the residual categories are necessarily low risk.

This is a dynamic space, considerations are given as to the balance between a strict liability approach driven by the protection of the people/society and openness to progress, creativity, and innovation in a sector where there is constant and rapid development.

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?

Australia's capacity to generate exports of AI-enabled technologies should not prevent the Department from upholding the same ethical and technical standards that it would insist upon for domestically sold products and services. That is not to say that Australia should impose a blanket ban on the development of high-risk AI technologies. Targeted banning of certain applications of AI that are incompatible with Australia's values on a case-by-case basis would be a more effective approach. A more long-term mitigation strategy would be to inculcate the next generation of developers by supporting the teaching of ethics at universities and incentivising the development of ethical codes of conduct within Australian industry. Furthermore, coordination with key allies and partners (including the US and UK) would limit the potential impact on Australian exporters by shifting a larger section of the market toward our ethical standards. Finally, the Department of Foreign Affairs and Trade could limit the impact of any such bans by working with neighbouring states to spread these standards throughout the region, limiting the number of other exporters willing to develop social scoring algorithms.