Safe and responsible AI in Australia

A response to the Discussion paper

This document contains a response to the questions:

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?

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

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?

Unintended behaviours or outcomes from AI pose a general yet significant risk due to the complexity of certain AI algorithms and their interaction with an AI environment. These unintended behaviour or outcomes can range from minor errors or inaccuracies to potentially harmful consequences that impact the safety[1] and well-being of society[2]. For example, the increasing number of reported AI incidents has shown a trend of generative AI tools being used to create misinformation[3] that is accessible to the world. Misinformation existed before generative AI was available at scale, the rapid uptake of generative AI has amplified the likelihood of public manipulation through misinformation, with the potential to impact society and global security at large.

This is a current example, and as we are still in the early stages of AI development and deployment, we are already witnessing the widespread impact of misinformation capturing public attention through social media platforms. Imagine the possible unintended outcomes if we don't have checks in place to manage this risk in the future.

Acknowledging the unintended behaviours or outcomes from AI, relating but not limited to:

- the use of personal identifiable information, which could be mitigated through strategies to create greater transparency, as indicated by the result from a recent Privacy Act Review report;
- creating bias and discrimination in Auto-Decision Making, which could be mitigated via existing Australia's anti-discrimination laws.
- online safety concerns, which are being addressed through the work of eSafety Commissioner.

Currently, there is no single comprehensive AI regulation to manage the general risks of AI, where these risks are widespread and common across different AI technologies and applications. Such risks occur throughout the AI development lifecycle including unintended behaviours or outcomes from AI.

Major technology companies, like Google, Microsoft and OpenAI with the capability to conduct research, build and deploy AI products, or features in their existing technology, are aware of

[1] Incident 545: Chatbot Tessa gives unauthorized diet advice to users seeking help for eating disorders https://incidentdatabase.ai/cite/545/#r3147, accessed July 2023

[2] Australian Human Rights Association, World Economic Forum, Artificial Intelligence: governance and leadership White paper, January 2019 https://tech.humanrights.gov.au/sites/default/files/2021-05/AHRC_RightsTech_2019_AI_whitepaper.pdf, accessed July 2023

[3] Janet Schwartz & Khoa Lam. AI Incident Roundup – May & June 2023, https://incidentdatabase.ai/blog/incident-report-2023-may-june/, posted 2023-07-18, accessed July 2023

the risks[4]. Recently they established a '*Frontier Model Forum, an industry body focused on ensuring safe and responsible development of frontier AI models*'[5] while simultaneously participating in a highly competitive race to dominate the market. AI regulation may create a balance to win in the development and/or deployment of AI without compromising safety and security, by placing legal obligations on all organisations to develop and deploy safe and secure AI products.

Attachment C: Possible elements of a draft risk-based approach could be further enhanced to consider a reporting obligation that mandates:

- Identification and central registration of AI models or products that meet a specific risk level, similar to the proposed European Union AI Act risk level requirements;
- Disclosure of unintended outcome that meet defined criteria compromising the safety and security of society, similarly to the Critical infrastructure reporting and compliance mandate in the *Security of Critical Infrastructure Act 2018* (SOCI Act).

Implementing a mandated comprehensive AI regulation using a risk-based approach will set clear expectations for both public and private organisation and holding organisations accountable, across the supply chain on research, design, develop and deployment of AI-enabled technology or AI products that protect the safety and well-being of society and the interest of national security, aligning efforts with other global leaders.

[4] Dan Milmo https://www.theguardian.com/technology/2023/may/04/us-announces-measures-to-address-risk-of-artificial-intelligence-arms-race, 5 May 2023, accessed July 2023

[5] https://blog.google/outreach-initiatives/public-policy/google-microsoft-openai-anthropic-frontier-model-forum/?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=a-new-ai-image-generator-is-in-town, posted 26 July, accessed July 2023.