Submission to the Proposals paper for introducing mandatory guardrails for AI in high-risk settings

September 2024

percapita
FIGHTING INEQUALITY IN AUSTRALIA

The Centre of the Public Square
A Per Capita initiative

SUBMISSION TO THE PROPOSALS PAPER FOR INTRODUCING MANDATORY GUARDRAILS FOR AI IN HIGH-RISK SETTINGS

The Centre of the Public Square (CPS) at Per Capita welcomes the opportunity to provide a submission in response to the proposals paper for introducing mandatory guardrails for AI in high-risk settings.

Per Capita is an independent think tank, dedicated to fighting inequality in Australia. We work to build a new vision for Australia, based on fairness, shared prosperity, and social justice. The Centre of the Public Square works to create equity and fairness for Australians online by holding technology companies to account and building better models of citizen collaboration through new methodologies and alternate technologies for the Australian public.

This submission outlines our recommendations on how to ensure AI is developed and regulated for the benefit of all Australians, and not just for the benefit of the large, multinational companies which produce AI models and products. In particular we will address the following questions from the proposals paper:

- Do the proposed principles, supported by examples, give enough clarity and certainty on high-risk AI settings and high-risk AI models? Is a more detailed approach, with a list of illustrative uses, needed?
- Do the proposed mandatory guardrails distribute responsibility across the AI supply chain and throughout the AI lifecycle appropriately? For example are the requirements assigned to developers and deployers appropriate?
- Do you have suggestions for reducing the regulatory burden on small-to-medium sizes businesses applying guardrails?
- Which legislative option do you feel will best address the use of AI in high-risk settings? What opportunities should the government take into account in considering each approach?
- Which regulatory option/s will best ensure the guardrails for high-risk AI can adapt and respond to step-changes in technology?

Summary

Governments around the world are now recognising both the significant potential opportunities as well as risks associated with artificial intelligence (AI). As the Australian government seeks to regulate AI, it's important that developments are for the benefit of the majority of Australians, and not just for the large technology companies that are creating AI models and products.
We recommend that:

1) **Australia establishes an AI Commissioner,** who is able to take a whole-of-economy approach to regulate AI under a new Australian AI Act

2) **Details of 'high-risk' settings for AI are specified and clarified in a list**, listing what these specific activities are, so that there is no confusion or uncertainty

3) **A similar definitional list is adopted for 'medium-risk' and 'low-risk' activities** to minimise the potential of any activities being misattributed or miscategorised, and address the approach for other related AI issues such as privacy or misinformation.

4) **That the position for ex-post facto applications and liability of these proposals are clarified**, given many significant AI models and products are already in market, and continuing to be updated at a rapid rate.

Australian AI Commissioner

Australia should establish an AI Commissioner who is able to take a whole-of-economy approach to regulate AI under a new Australian AI Act.

Of the approaches proposed - domain specific, framework, or whole-of-economy, it is clear that a whole-of-economy approach would be best suited in capturing the relevant considerations and issues that AI presents.

Key to this is the speed of technological development and the potential step-changes AI facilitates.[1] The proposals paper describes well how AI is different from other technologies – including its potential for autonomy, its adaptability and ability to learn, its speed and ability to scale, and the black box nature of models and outputs.

Large technology companies' commitments and investments for AI show no sign of slowing down, despite fears that returns won't materialise.[2] Anything other than a dedicated authority whose sole focus is in managing and keeping abreast of developments would prove too slow and ineffectual to react to those changes.

An Australian AI Commissioner could investigate and prosecute any harms or breaches done by AI companies, especially as it relates to medium to high-risk activities identified.

The Commission could also engage technology companies in conversations around minimising harms and developing risk mitigation strategies for AI activities ongoing.

The Commission should also champion local capability by negotiating with large, foreign AI companies on any value exchange, compensation, industry-wide partnerships, collective bargaining agreements and funding based on revenue generated from Australia.

The AI Commissioner should also work closely with other regulators and government departments, particularly on other legislation where AI has material impacts, like privacy and misinformation. The Commission could act as a specialist technical resource for other government offices, providing advice, submissions, research and insights.

[1] Gruetzemacher, R & Whittlestone, J. 2022, The transformative potential of artificial intelligence, *Futures Journal,* Volume 135.
[2] Dang, S. 2024, *AI spending worries cast gloom over Alphabet, Microsoft*, https://www.reuters.com/technology/ai-spending-worries-cast-gloom-over-alphabet-microsoft-2024-04-25/

Clarify 'high-risk' activities in a list

Given the potential for wide application of AI across many industries, professions and practices, a list-based approach in defining 'high-risk' activities would provide clarity and transparency for regulators, companies and the general public.

There have already been several principles, mission statements, and position documents developed across different levels of government, the public service and technology companies. And while these can be useful 'statements of intent' further clarity and more detailed definitions are needed to progress effective regulation and legislation.

These lists don't need to be completely exhaustive, but provide enough definitions and examples to cover most use-cases. They also need to be regularly updated given the speed of development with AI technologies. How these are updated and the process of definition should be made transparent.

High-risk use cases defined in the proposals report from other countries is a useful place to start. But even these broad descriptions already need further clarification.

For example, one domain area that already need unpacking is the 'Administration of justice and democratic processes'. The general description states that "with regards to democratic processes, may include any system which can influence the voting behaviour of individuals or the outcome of an election or democratic process".[3]

This could apply to the whole news industry and media companies, which play a critical role in influencing voting behaviours. In other words, even a detailed list of domain areas already require clarifications and would need use cases, case studies and further instructions.

Support tools, like the free to use EU AI Act Compliance checker or risk assessment documents could assist companies and the public in identifying and defining the risk thresholds of relevant AI products and services.[4]

Medium-risk and other issues

There are currently no mentions of 'medium-risk' issues in the proposals paper. While care has been given in acknowledging 'low-risk' activities – mostly to ensure that they are not unintentionally captured in the stricter 'high-risk' categories, 'medium-risk' settings have not yet been acknowledged.

This could potentially be the most contentious and the biggest grey area among the risk categories.

The EU AI Act refers to 'medium-risk' as 'limited risk' systems, which are AI systems with extra transparency requirements. These include chatbots and apply to the majority of current generative AI products and uses.

---

[3] Department of Industry, Science and Resources, 2024, *Proposals paper for introducing mandatory guardrails for AI in high-risk settings*
[4] EU Artificial Intelligence Act, *EU AI Act Compliance Checker*, accessed Sept 2024, https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker/

Technically, the EU definition could apply to deepfakes, and would only require a label or watermark stating that piece of content is AI-generated. This begs the question of whether a label would be enough to cover the potential damages of a deepfake video, for example during an election campaign. In the EU definition knowledge or consent is the differentiator between a limited or medium risk, and high risk, or "subliminal manipulation".[5]

You could argue that even if a deepfake video is labelled, there are those who would still be fooled or would ignore the warning. Moreover, there are questions about whether we would find it acceptable that we allow a flood of deepfake videos to pollute our information ecosystem, as long as they are labelled accordingly.

This is but one example of the tension between what could be classified as 'high-risk' or 'medium-risk' that needs to be further interrogated.

There should also be details and acknowledgment of any activities that may fall under separate jurisdictions, using not necessarily the same list mechanism, but should certainly be defined. There should be clarification of the approach and developments to regulating those activities, such as AI activities that may fall under the Privacy Act or the Mis/Disinformation Bill and how both agencies address these.

Ex-post facto applications and liability

The guardrails proposed are a welcome development in ensuring AI systems are as safe as possible. The approach of trying to establish guardrails upstream, as early in the development cycle as possible, is also to be commended.

However, there are already several major AI models and products in train and released in the market, flowing downstream to users. It's critical that these guardrails are applied to these models as well, and newer, better, compliant versions are released to replace any existing ones which breach any guardrails.

For example, guardrail 3 states that systems have appropriate data governance and ensure data quality. It states that data should be legally obtained, that data should not contain illegal and harmful material, and that data is fit for purpose.

Even briefly applying this guardrail would see many current AI model owners, like OpenAI, Meta and Alphabet fail given that data they collected to train their models were not fit for purpose and users were not given any choice in the matter[6], and can contain harmful material.[7]

It should be clarified then that these companies should be liable for any harms that result in these non-compliant models and products. Further, the enforcement mechanisms for breaches need to be substantial enough for large technology companies with vast resources to take seriously.

---

[5] EU Artificial Intelligence Act, *High-level summary of the Act,* accessed Sept 2024, https://artificialintelligenceact.eu/high-level-summary/

[6] Evans, J. 2024, *Facebook admits to scraping every Australian adult user's public photos and posts to train AI, with no opt-out option*, https://www.abc.net.au/news/2024-09-11/facebook-scraping-photos-data-no-opt-out/104336170

[7] David, E, 2023, *AI image training dataset found to include child sexual abuse imagery*, https://www.theverge.com/2023/12/20/24009418/generative-ai-image-laion-csam-google-stability-stanford

Breaches of the EU AI Act can incur a penalty of up to 7% of annual turnover.[8] The Australian AI Act should consider similarly material penalties.

It would also be worth specifying a particular clause for very large online platforms who disproportionately own and develop AI foundation models, in contrast with the companies who deploy those models.

Given that AI models from companies like OpenAI, Alphabet, Meta, Microsoft and NVDIA disproportionately cover the majority of the AI market and AI products, these companies should be put under extra scrutiny given their outsized influence in the AI space.[9]

This would also mean that any AI use downstream including by local small-to-medium businesses will have less regulatory pressure and compliance requirements to consider, if the scrutiny and safety requirements are strictly enforced and are upheld from the source/the originators of AI models and AI products.

Conclusion

The work of regulating AI is underway. With significant international references to draw from, the government should be well placed to develop an Australian AI Act locally.

An Australian AI Commissioner will ensure that a whole-of-economy approach can cover the needs of the Australian public and local businesses.

An Australian AI Act should also clarify definitions of high, medium and low-risk activities to prevent confusion and uncertainty.

Any guardrails developed should also apply to existing AI models and products, and ensure the large technology companies who own them are held to account and comply with our laws.

[8] European Commission, *European Artificial Intelligence Act comes into force,* https://ec.europa.eu/commission/presscorner/detail/en/IP_24_4123
[9] Morris et. al. 2024, *Big Tech's AI dealmaking needs 'urgent' scrutiny, says US antitrust enforcer,* https://www.ft.com/content/97b45759-36e0-4f5b-9c6a-ae0580f9a29b