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

SUBMISSION TO THE DEPARTMENT OF INDUSTRY, SCIENCE AND RESOURCES

October 2024

INTRODUCTION

1. We thank the Department of Industry, Science and Resources (**Department**) for the opportunity to comment on its proposals for regulating artificial intelligence (**AI**).

2. ANZ uses AI in various ways to assist and protect our customers. In using AI, our priority is ensuring that our customers are treated fairly, and that their information is secure.

3. We support the Department's objective of maximising the benefits of AI to the Australian community while minimising harms. It is important that regulatory settings support the safe and responsible use of AI, but do not introduce barriers to the development and use of AI to protect customers, innovate, improve products and services, and enhance productivity.

4. Accordingly, we support the Department's proposed targeting of high-risk AI. We also support the Department's stated focus on international interoperability.

5. To assist the Department to refine its proposals and determine the appropriate regulatory settings, we have set out some observations below on defining high-risk AI, the proposed mandatory guardrails, and how to incorporate the guardrails into law.

DETAILED POINTS

Defining high-risk AI

6. The Department's proposals paper proposes two broad categories of high-risk AI: high-risk known or foreseeable *use or application* of AI (defined by reference to proposed principles), and general-purpose AI (**GPAI**) models, where all the possible uses, risks and applications cannot be foreseen. The proposed principles for defining high-risk AI use are broad, with an intention of providing regulatory flexibility. The paper proposes to capture all GPAI models.

7. To achieve an appropriate balance between ensuring responsible AI use and encouraging its adoption, it is important that regulation is appropriately targeted at AI that creates an unacceptable risk to end users and the Australian public. It is also important, to help ensure that regulation can be efficiently and effectively applied by regulated entities, that the need for flexibility is appropriately balanced with that for regulatory certainty.

8. As proposed, the principles for defining high-risk AI *use* carry the risk of being applied inconsistently by regulated entities, and/or capturing uses that may not be high-risk and/or have a high customer or public utility. For example:

    - The proposed principle relating to the 'risk of adverse legal effects, defamation or similarly significant effects' could capture fraud, scam and financial crime detection mechanisms that analyse customer behaviour to identify patterns, deviations and

apply rules to minimise the impact of fraud. When fraud is detected, for example, these mechanisms can take actions that could have legal effects, such as declining a payment made by a perpetrator. These mechanisms are essential to help protect customers.

- The same principle, or the principle relating to 'adverse impacts to groups of individuals', could capture the use of AI to identify financial abuse or customer vulnerabilities, including the need for support following a natural disaster or upon falling into financial difficulty. These mechanisms are in place, or could be developed, to help protect and support customers.

- The proposed definition does not set out a test against which to assess risk, and the severity and extent of the adverse impacts. While regard must be had to these factors, it is left to regulated entities to determine how to weigh these risks, including how to balance likelihood with severity.

9. To the extent that high-risk AI use is regulated, we think that a list-based approach with more specifically defined captured uses would be preferable. This would better provide for regulatory certainty and consistent application, and would allow for regulation to be targeted at known high-risk uses. To help ensure flexibility and allow for 'future-proofing', a mechanism could be included to provide for the list to updated as appropriate.

10. A list-based approach is taken by the European Union in its *Artificial Intelligence Act* (**the EU Act**).[1] Adopting a list here could, therefore, help facilitate interoperability with the EU. The EU Act is nascent, and its effectiveness and impacts on AI adoption and innovation are, as yet, unknown. Where appropriate, though, alignment with the framing of the EU Act could help facilitate regulatory efficiency for regulated entities.[2]

11. We also think that the definition could, like the EU Act,[3] explicitly (and non-exhaustively) carve out uses that are not high-risk (whether the definition is in a principles or list-based form). This could include some of the matters included in the EU Act, for example when AI is intended to improve the result of a previously completed human activity, or when AI is intended to perform a preparatory task to an assessment relevant to a high-risk use case. It could also include uses, such as those described in paragraph 8, that have a clear customer or public benefit.

---

[1] See Annex III.
[2] We note, however, that interoperability with the US is important, given this is where many AI developers are based. The US does not have federal AI legislation.
[3] Article 6(3) https://artificialintelligenceact.eu/article/6/.

12. With respect to GPAI models, we think, to ensure that regulation is commensurate with risk, there should be a distinction between advanced, highly capable models that could present a high or systemic risk and others. This distinction could be based on the model's capabilities and/or the potential nature and scale of use. Regulation could be scaled to reflect these models' higher risk.[4]

Guardrails

13. The proposals paper sets out ten mandatory guardrails that would apply to high-risk AI. The principles would apply to both AI deployers and developers, and appear to apply to all forms of high-risk AI captured by the proposed definition. Like the proposed principles for defining high-risk AI use, the guardrails are framed broadly.

14. The guardrails reflect several factors we consider when dealing with AI, including fairness, transparency, contestability, and governance.

15. It is, though, important that the guardrails are sufficiently precise to enable regulated entities to understand and comply with their obligations. It is also important that the guardrails can be effectively implemented by entities and do not unintentionally unnecessarily restrict the adoption of AI.

16. In these respects, we make the following comments:

    - We think that the requirements of developers should be clearly distinguished from those of deployers, given the distinct roles each can play in mitigating AI risk. The proposals paper sets out at Attachment E, at a high level, what the guardrails may require of each of developers and deployers of AI. We think that the final form of the guardrails should specify the different obligations. We note that the Department's Voluntary AI Safety Standard does not clearly set out this distinction.

    - The guardrails do not contain any territorial nexus. It is not clear whether and how, for example, the guardrails would apply to developers or deployers of AI systems or GPAI models that are based outside of Australia. Guardrails should be workable for international organisations, and, to the extent possible, consistent with requirements in overseas jurisdictions. We note that if developers based or deploying AI in Australia face more stringent requirements than those based or deploying AI overseas, this could create barriers for AI adoption.

[4] We note that the EU Act has a test for classification of a GPAI model as having systemic risk at Article 51. These models are subject to additional requirements.

- Proposed Guardrail 5 currently requires enabling human control or intervention in an AI system to achieve meaningful human oversight. It is unclear how human control or intervention could be enabled for some automated decisions or machine learning processes where the AI's reasoning is not visible. It is also unclear how the control or intervention requirement can, in practice, be applied to GPAI models given that the model/algorithm operates independently of humans. We think the guardrail should, like the requirements at Articles 14 and 26(2) of the EU Act, be targeted at human oversight, rather than control or intervention. This is consistent with how the obligations are framed at Attachment E.

- The requirement to inform end-users in Proposed Guardrail 6 is broader than the obligations in Article 50 the EU Act, which apply to AI interactions and AI-generated content, but not specifically to AI decisions. Insofar as Proposed Guardrail 6 is proposed to apply to decisions, it should be consistent with and not extend beyond the final enacted requirements relating to automated decision-making that may be introduced to the Privacy Act by the *Privacy and Other Legislation Amendment Bill 2024.*[5] Ideally, the requirements to inform customers of AI use should all be addressed in the one regulatory instrument, and not duplicated.

- Proposed Guardrail 7 requires entities to establish processes to allow challenge the use of or outcomes produced by AI. We think these requirements should be limited to the AI systems and uses covered by Proposed Guardrail 6. It is not feasible to allow customers to challenge some uses of AI, such as when AI is used to improve efficiency of internal processes or as part of fraud and scam detection activity. We think this requirement should be able to be met through existing complaints mechanisms, rather than the establishment of new parallel processes.

- The transparency requirements of Proposed Guardrail 8, as they apply to developers, should not require the sharing of information (which may include intellectual property) by an organisation that has developed an AI solution for its own internal purposes.

- To help with international interoperability and promote efficiency, the requirement of proposed Guardrail 10 to undertake conformity assessments should, to the extent possible, align with requirements of conformity assessments of international jurisdictions, such as the EU.[6]

[5] This Bill is currently before Parliament.
[6] See, for example, Article 47 and Annex V of the EU Act.

17. We would also benefit from clarification as to how the guardrails would apply to processes involving multiple AI systems (such as multi-agent systems), and open source GPAI models.

How to incorporate guardrails

18. The proposals paper considers regulatory options to incorporate the proposed guardrails into law, including a domain-specific approach, a framework approach and a whole-of-economy approach.

19. The appropriate approach should seek to minimise regulatory duplication while ensuring that, to the extent it is appropriate, there is a consistent regulatory approach to AI as it is used across the economy and in different contexts. It must also be able to adequately capture and apply to developers, who may not be covered by other regulatory regimes in the same way as deployers.

20. Financial institutions engage with several regulatory regimes and regulators. These regimes address, or have the capacity to address, many of the risks of AI. To help promote regulatory efficiency, a sectoral approach to incorporating the guardrails, whereby existing financial services regulation could be utilised, would be preferable.

21. Whichever approach is taken, regulatory regimes should take a complementary approach to AI. The requirements should be consistent and, to the extent possible, not overlap. For example, as indicated above, the approach taken to automated decision-making in the Privacy Act should be consistent with, and not duplicate, the requirements of the guardrails.

ENDS