RESERVE BANK OF AUSTRALIA

Response to the Proposal Paper for Introducing Mandatory Guardrails for AI in High-risk Settings

Department of Industry, Science and Resources

October 2024

The Reserve Bank of Australia appreciates the opportunity to provide feedback on the proposals paper, 'Introducing Mandatory Guardrails for AI in High-risk Settings'. We are not making specific recommendations. However, we would like to raise some concerns and seek clarity around certain aspects of the paper.

1. Defining high-risk AI

The proposals paper attempts to clarify the concept of general-purpose AI (GPAI), including on page 8 where it states:

> *General-purpose AI (GPAI) models (that is, AI systems that could be used for a wide range of purposes) are the next evolution of AI. GPAI models, such as GPT-n, DALL-E, and Sora, can now generate 'human-like' text, images, and videos based on simple user prompts.*

This description lacks specificity. Further, the definition of GPAI on page 9 as '*an AI model that is capable of being used, or capable of being adapted for use, for a variety of purposes, both for direct use as well as for integration in other systems*' does not clearly distinguish it from narrow AI systems. Narrow AI is defined in the paper as:

> *[A] type of AI system or model that is focused on defined tasks and uses to address a specific problem. Unlike GPAI models, these types of AI systems cannot be used for a broader range of problems without being re-designed.*

The distinction between 'adapting' a model and 're-designing' it is ambiguous. For example, a logistic regression model designed to classify flowers – which is a narrow AI application – can be re-designed to classify trees. This re-design process could be interpreted as adaptation, thus falling under the current definition of GPAI.

Given this ambiguity, we suggest dedicating a section of the paper to clearly define GPAI and narrow AI. As it stands, the current definition risks unintentionally categorising low-risk use cases as GPAI.

In addition, the proposed principles-based approach to defining high-risk AI seems overly broad, potentially capturing low-risk applications. This could deter the use of models to produce rigorous assessments of public policies and/or deter cautious entities from investing in AI. For example, in an economics or finance setting, it is not clear if a machine learning model used as an input (among many) when assessing the impact of a public policy or producing a forecast would be classified as high risk. Without greater clarity in the definitions of high-risk AI, broad regulations could inadvertently encompass long-standing applications of AI and machine learning models used as inputs (among many) for analytical purposes, leading to unintended consequences and inefficiencies in fulfilling our charter functions and thereby serving the public.

There is also ambiguity about the role an AI model must play in decision-making to be considered high risk. For instance, if generative AI is used to produce a variable or indicator that constitutes just one of many inputs into a policy deliberation – including human judgement – would the use of AI in this context be considered high risk? Clearer guidelines on the significance of AI in decision-making processes are needed.

2. Guardrails ensuring testing, transparency and accountability of AI

A clear risk assessment methodology with examples of high-, medium-, and low-risk categories could make the guardrails more meaningful. The controls framework should align with and expand on existing guidance, such as the Australian Cyber Security Centre's guidelines for deploying AI systems securely.[1] This approach would allow for independent updates and ensure that existing investments in controls could be reused and expanded.

Establishing an AI provider certification framework would be helpful for independent testing and verification against a standard, indicating that AI technologies are safe to use. An example is the hosting certification framework for cloud providers managed by the Australian government.

There also needs to be clear guidance on the rules that apply to consumers, developers and platforms.

Further, there should be specific provisions addressing data leakage, data usage and attestation to ensure data security and privacy. Additionally, AI-generated content could be clearly labelled, similar to food labelling (e.g. 'may contain nuts'), to inform consumers that the content was generated using AI.

3. Regulatory options to mandate guardrails

We understand that the proposals paper outlines three regulatory options: adapting existing regulatory frameworks, introducing framework legislation and creating a new AI-specific Act. We do not prefer one option over the others. However, we would value further clarity on how each option would be implemented and their potential impact on existing regulatory frameworks and industry practices.

In conclusion, in our view refining the proposals in the paper to address the concerns raised above and to provide clarity on definitions, guardrails and regulatory options will help create a balanced regulatory environment that promotes innovation while ensuring the safe and responsible use of AI. We appreciate the opportunity to provide feedback and look forward to continued engagement on this important issue.

Reserve Bank of Australia
3 October 2024

[1] Australian Cyber Security Centre (2024), 'Deploying AI Systems Securely', Report, 16 April.