TCA Summary:
Proposals Paper for Introducing Mandatory Guardrails for AI in High-Risk Settings and Voluntary AI Standard

Key points

- On Thursday 5 September, the Federal Government released a proposals paper for Introducing Mandatory Guardrails for AI in High-Risk Settings, meeting its earlier commitment in the Interim Response to Safe and Responsible AI.
- The proposals paper contains mandatory guardrails for AI, and is accompanied by the release of a Voluntary AI Safety Standard. This is designed to support and promote best-practice governance to help businesses start to adopt AI in a safe and responsible way.
- The proposals paper seeks feedback on: (i) proposed definitions of 'high-risk' AI uses and principles, (ii) the 10 proposed mandatory guardrails, and (iii) three options to consider adopting the guardrails in Australia's regulatory regime (existing regulatory frameworks, new framework legislation, an AI-specific Act).

TCA Response

- TCA welcomes the Government's progress on safe and responsible AI in Australia. We support the release of the voluntary standard and the proposals paper for mandatory safeguards for high-risk AI systems.
- These are crucial steps in building an appropriately balanced regulatory strategy for AI governance in Australia to support confidence and trust in AI technologies.
- We are pleased to see the Government adopt a risk-based approach, consistency between the standards, and consider a range of options to adopt mandatory standards.
- It is important that Australia's regulatory approach is well balanced to encourage innovation and be sufficiently flexible to accommodate new and emerging techniques for responsible AI oversight.
- We are supportive of options that leverage the expertise of well-established regulators that is well coordinated. We continue to encourage the Government to invest in Australian AI capability as a nationally significant critical technology.
- There are open questions with regard to the mandatory guardrails 'principles', thresholds, and how they will be applied to organisations, as well as whether these are self-assessed or otherwise determined. Questions also arise surrounding enforcement and compliance for organisations that are using AI models in high-risk settings.

Proposals paper for mandatory guardrails for AI in high-risk settings

- The proposed mandatory guardrails are measures that would require developers and deployers of AI in high-risk settings to take specific steps across the AI lifecycle.
- The paper proposes a principles-based approach to defining high-risk AI with 'known or foreseeable uses' and considers whether a list-based definition of 'high-risk' should be adopted, similar to other jurisdictions.

<u>Proposed principles to determine whether an AI use-case is 'high-risk'</u>

a) The risk of adverse impacts to an individual's rights recognised in Australian human rights law without justification, in addition to Australia's international human rights law obligations

b) The risk of adverse impacts to an individual's physical or mental health or safety

c) The risk of adverse legal effects, defamation or similarly significant effects on an individual

d) The risk of adverse impacts to groups of individuals or collective rights of cultural groups

e) The risk of adverse impacts to the broader Australian economy, society, environment and rule of law

f) The severity and extent of those adverse impacts outlined in principles (a) to (e) above.

As expected, there are 10 proposed mandatory guardrails that could apply across the AI supply chain:

1. Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance.
2. Establish and implement a risk management process to identify and mitigate risks.
3. Protect AI systems and implement data governance measures to manage data quality and provenance.
4. Test AI models and systems to evaluate model performance and monitor the system once deployed.
5. Enable human control or intervention in an AI system to achieve meaningful human oversight.
6. Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content.
7. Establish processes for people impacted by AI systems to challenge use or outcomes.
8. Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks.
9. Keep and maintain records to allow third parties to assess compliance with guardrails.
10. Undertake conformity assessments to demonstrate and certify compliance with the guardrails.

The Government proposes three regulatory options to introduce mandatory guardrails

1. Adopting the guardrails within existing regulatory frameworks as needed
2. Introducing new framework legislation to adapt existing regulatory frameworks across the economy
3. Introducing a new cross-economy AI-specific Act (for example, an Australian AI Act).

Voluntary AI Standard

- The Standard gives practical guidance to all Australian organisations on how to safely and responsibly use and innovate with AI. This Standard was developed by the National AI Centre (NAIC).
- The version one of the Standard is aimed at deployers of AI systems working with third parties (developers).
- The Standard consists of ten guardrails intended to be consistent with the *mandatory guardrails for high-risk AI systems* that outline expectations for organisations throughout the AI supply chain.

1. Establish, implement, and publish an accountability process including governance, internal capability and a strategy for regulatory compliance.
2. Establish and implement a risk management process to identify and mitigate risks.
3. Protect AI systems, and implement data governance measures to manage data quality and provenance.
4. Test AI models and systems to evaluate model performance and monitor the system once deployed.
5. Enable human control or intervention in an AI system to achieve meaningful human oversight across the lifecycle.
6. Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content.
7. Establish processes for people impacted by AI systems to challenge use or outcomes.
8. Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks.
9. Keep and maintain records to allow third parties to assess compliance with guardrails.
10. Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness.

- The Standard draws on and is intended to aligned with a range of international standards. Most important is the leading international standard on AI Management Systems, AS ISO/IEC 42001:2023 and the US NIST AI RMF 1.0.

Next steps

- TCA will be providing two briefing and information sessions on the proposals paper on mandatory guardrails for high-risk AI on **Friday 6th September at 3.30pm** and **Monday 9th September at 10am** via Microsoft Teams. You may choose to attend either of these.
- We will reconvene our Data and AI Working Group on **Tuesday 10th September at 2pm** to seek feedback on the proposals paper once members have had the opportunity to digest materials.
- We will work with members over the next four (4) weeks to provide a TCA submission to the consultation which closes on Friday, 4 October.
- If you are not part of the Data & AI working group and would like to be contribute to this work, or would like to provide member feedback out of session, please reach out to Erika (erika@techcouncil.com.au)