Response to the 'Proposals Paper for introducing mandatory guardrails for AI in high-risk settings' by the ARC Centre of Excellence for the Digital Child

Submission made by the **Australian Research Council (ARC) Centre of Excellence for the Digital Child**, prepared by Professor Tama Leaver (Curtin University).

About The ARC Centre of Excellence for the Digital Child

The ARC Centre of Excellence for the Digital Child is shaping an environment with children, families, and communities so they can navigate their own digital worlds. We know that children's lived experiences are rapidly changing, and that every childhood is now fundamentally digital. Our mission is to create positive digital childhoods for every child in Australia.

The Digital Child is a national research centre comprised of over 200 researchers from across Australia in addition to national and international partners. Our research is focused on: *healthy digital lives*, understanding how digital technology intersects children's lived experiences and providing guidance to families, educators, and policymakers as they navigate this space; *educational empowerment*, equipping children with the skills they need to live their best digital lives; and *safe digital spaces*, making online engagement safer while promoting healthy digital relationships.

The Digital Child is funded by the Australian Research Council. Our research includes the world-first Australian Children of the Digital Age longitudinal study, which is tracking digital engagement of more than 3,000 Australian families for four years.

General Response

We commend the development of the Mandatory Guardrails approach, and the resourcing and speed at which this process is happening. For the most part, the proposed principles and guardrails do extremely important work and will go a long way toward protecting Australians using AI tools. However, looking through the lens of Australian children, there is some specific language required to delineate the potentially differentiated experience of children using AI tools. This language should ensure that any transparency in how AI tools operate should be accessible in age-appropriate language, meaning that if, for example, an AI tool is being useful commercially or in educational settings by 8 or 13 year olds, then descriptions of how it operates, what data it collects, and so forth, should be understandable in terms those child end-users can readily understand and base meaningful action on (such as the action to not use the tools).

Defining high-risk AI

The principles are doing a lot of work to try and flag who is exposed to risk, but does not specifically flag the question of risks to children. Explicitly flagging risk in age-appropriate ways which notes the different experiences of children should be included in the proposed principles. It's worth noting that as they stand, almost all educational uses in the K-12 sector would seemingly be high-risk. This is probably appropriate.

Explicit mention of Indigenous Australians is needed, especially in terms of addressing culturally specific uses of data which may impact Indigenous peoples differently given their cultural traditions.

We recommend an additional axis of risk be added: "The risk of adverse impacts on children in terms of age-appropriate uses and understandings".

Guardrails ensuring testing, transparency and accountability of AI

Any AI tool operating in Australia that can generate content about Indigenous Peoples should have a specific guardrail preventing the creation on outputs that might compete with intellectual property or outputs by Indigenous Australians. It's very hard to know in advance. Hence the guardrails should be reviewed annually to begin with. We prefer the notion of a new cross-economy AI act. A specific AI act could be more readily updated as needs be, which may well be needed as the shape and scope of AI are altering at a significant pace. Being wrapped in other legislation may slow any changes being made.

The fact that AI tools pretend to be agentic (ie human-like decision makers) is especially difficult for younger children to comprehend as false, especially where voice interfaces are concerned. Ensuring AI tools and agents do not fool (purposefully or otherwise) young people into disclosing personal information, or forming parasocial attachments to AI tools, is important.

The guardrails as they stand do not specifically acknowledge the different experiences of AI tools for children (anyone under 18). Ensuring any explanation of AI operations are understandable to users is vital. Embedding this in the guardrails could be achieved by inserting the words 'age appropriate' into (6), so it now reads **"Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content in an age appropriate manner."** This would ensure AI tools used by children can be understood in language accessible to children.

Clarity Between Deployers and Developers

A line is draw between the companies creating AI (developers) and those using AI as a tool (deployers), but there is an awkward grey area in some of the language. It is not clear, for example, if additional training is done by a deployer, do they then end up being classified as a developer? This is especially important given the roll out of 'MyGPTs' and other services which encourage users to focus AI tools onto a smaller more specific data corpus.