Submission No 15

INQUIRY INTO ARTIFICIAL INTELLIGENCE (AI) IN NEW SOUTH WALES

Name: Ms Naomi Murn

Date Received: 16 October 2023

I appreciate the opportunity to make a submission and share my views about the potential impact of AI on NSW.

As a lifelong NSW resident with an interest in tech and automating mundane tasks, I have been reflecting on AI and the role it has in my daily life, and how that role may change in the future.

I appreciate NSW's proactive approach in respect of the AI policy and assurance framework, in particular compared to other jurisdictions. It is evident that NSW understands that the impact of AI must be deeply thought about - there is potential for it to transform our society as we know it. I want the NSW government to make informed and timely decisions to ensure we remain at the forefront of addressing emerging challenges and risks from AI. I have set out two recommendations below which will assist with this.

Australia as a leader in AI labs - and how NSW can assist

Terms of reference 1.(k) asked after measures other jurisdictions, both international and domestic, are adopting in regard to the adaption to and regulation of AI.

One global developments which stands out is the creation of "national laboratories" to enable technical tests on AI models, provide technical reports and provide ongoing monitoring and assurance. As examples, Singapore has established the AI Verify Foundation, the EU has created a Centre for Algorithmic Transparency, the UK has a Foundation Model Taskforce and the Tony Blair Institute for Global Change has proposed that the UK create "Sentinel" with a similar goal.

Without a comparable lab in Australia or in the region, deploying trusted and safe AI in Australia might become impossible as capability and capacity increases.

NSW is well positioned to collaborate with other jurisdictions to create or support a national laboratory for AI safety, modelled on international best practices.

This approach is exciting because there's international best practice to follow, and NSW could use the laboratory to ensure the AI products it uses are safe and can be subject to effective ongoing monitoring and assurance.

Perhaps even more importantly, we are already seeing various kinds of dangerous and risky AIs. If we had a trusted national laboratory, it could assess AI products before they go to market. In the same way we don't let cars on our roads without them going through safety tests, a lab like this would allow us to block AIs until they've passed appropriate safety tests.

AI assurance framework

NSW has made a positive move in developing a transparent AI assurance framework. While the framework is an excellent start, there is room for immediate improvement.

One specific concern, most obviously expressed on page 13 of the assurance framework, assets that "the key factor that determines risk is how the AI system is used". At best this is misleading, and at worst, it's dangerous. This claim is likely to be false.

While the **use** of the system is a relevant factor, NSW should urgently pivot to focusing on **the risk of the system itself**, in addition to its particular use case. That is, NSW's Risk Assessment should develop a list of features that might make an AI more or less risky. For instance, systems that are more strictly a "black box", more likely to hallucinate, can act more autonomously, are less aligned with our values, or are frontier models with

cutting-edge capability and capacity should be considered more risky. On the other hand, systems that are transparent, human-interpretable, have been reviewed by world-leading AI safety labs, can be subject to practical ongoing review and are reliably controllable should be considered less risky.

We are starting to see even today that unpredictable systems can do dangerous things and cause harm even when the use case is only entertainment. As AIs gain more capabilities and become more autonomous, what will matter is how safe the AI itself is, not just our intentions of how we mean to use it.

In this context, I'd draw NSW's attention to Anthropic's recently published Responsible Scaling Policy. The policy details "AI Safety Levels (ASLs)" and sorts models from ASL-1 to ASL-4+. This is broadly similar to the Biosafety Levels (BSL) that are currently used to regulate labs that work with infectious diseases, including in NSW. NSW could overlay this system on its current risk assessment model to improve its performance.

Final comments

Overall, we know that we're on track for AI technology that continues to accelerate and transform our society. I hope that NSW continues to frequently re-think how best to configure its AI policies, adapt to emerging evidence, and encourage the other governments of Australia to do the same. We don't know today if AI is trending to make things very good, or very bad. What we do know is that we need vigilant governments that are watching these trends and are ready to act.