Submission No 22

INQUIRY INTO ARTIFICIAL INTELLIGENCE (AI) IN NEW SOUTH WALES

Name:	Name suppressed
Date Received:	20 October 2023

Thank you for the opportunity to make a submission and share my views about how AI might impact NSW, including the risks and challenges it presents.

As a younger member of the NSW community, I've been thinking a lot about AI and how it is already affecting my work and my community, and the impacts it could have in the future.

When I read NSW's AI policy and assurance framework, I appreciated that the NSW government has been ahead of other jurisdictions in understanding the pace of change of AI and its transformative nature. I want government to make the right calls in a timely way, and stay ahead of emerging issues and risks from AI. The recommendations I provide below hope to contribute to that.

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 of the global developments that is most promising is the creation of "national laboratories" to enable technical tests on AI models, provide technical reports and provide ongoing monitoring and assurance. 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 similar 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.

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, is that "the key factor that determines risk is how the AI system is used". At best this is misleading. It's probably not true. In the medium term, it's dangerous.

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 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.

Finally, I understand that there is a range of views about AI. Some people think that it could be an existential risk, and others think that it will solve all our problems. What I think everyone agrees on is that we need more investment in AI safety research. If existential risk is real, AI safety research could save humans from extinction. Even if existential risk is not real, AI safety research is going to avoid a range of harms and ensure AI is able to understand our intents and operate efficiently and effectively. In that context, anything NSW can do to foster research into those issues is going to make a positive difference.