Submission No 13

INQUIRY INTO ARTIFICIAL INTELLIGENCE (AI) IN NEW SOUTH WALES

Name: Name suppressed

Date Received: 16 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.

We need to urgently build state-based and national regulatory schemes for banning AI applications that are dangerous. We need to ensure dangerous AIs are not deployed to NSW, and hopefully disincentivise them from being developed in the first place.

The AI governance conversation often talks about the balance between regulation and social or economic benefit. I think this is not always the best way to think about the question. Some AIs are so dangerous, or offer so little benefit, that they just have no place in our society.

The obvious present example of this is "undress AIs" or "deep nudes". These are generative AI tools that allow their users to upload a normal picture of a person, select some parameters, and return a picture of that person naked. Just recently we saw news of a group of schoolboys in Spain using AI tools to generate fake images of dozens of schoolgirls from their small town.

This technology offers no pro-social use (perhaps outside of narrow research or law enforcement situations) and empowers harassment, fraud, the invasion of privacy and child abuse.

I don't think the banning of these products is a complex issue. If we don't act now, what has already happened in Spain is bound to happen here soon. We don't let people go around NSW with guns or spikes on their cars, and so we shouldn't let people in NSW have obviously dangerous AI products.

Building a regulatory framework that allows products like undress AIs to be banned will have lasting benefits. Given the pace of change, we don't know what other dangerous AI products could be just around the corner. I've read about medical AIs being used to make bioweapons, coding AIs being used to run round-the-clock cyber attacks and people making automated AIs and setting them loose on the internet with the only goal of making money. We don't know where this is going to go, and when capability will develop enough for these to move from academic curiosities to existential risks. We need to build the frameworks that give us ways to respond now so we can use them to address today's threats and be ready for the threats of tomorrow.

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.

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.