Submission No 6

INQUIRY INTO ARTIFICIAL INTELLIGENCE (AI) IN NEW SOUTH WALES

Name: Mr Leosha Trushin

Date Received: 23 September 2023

Hello NSW Parliament, I am Leosha Trushin, a Bachelor's student at ANU who has engaged with mathematics, computer science and AI courses and activities both inside and outside of university. The risk of extinction from potentially near-future superintelligent AI is deeply concerning and important to me. I believe that current developments in AI and the incentive structure in the field is leading towards highly capable, but naively hopeful and badly tested AI systems, and it is an existential risk to create something highly capable of doing things in the world without rigorously ensuring it will do what we want it to.

I'm glad that NSW has taken a forward-leaning approach to AI and is continuing to work to stay ahead of the fast-moving and transformative technology.

I think much of the conversation in Australia underestimates just how significant AI is – including just how bad it could be if it goes wrong. I think our governments have a responsibility to be aware of the real risks of AI and capably act to manage those risks. A glaring example of this was the Bing AI, which was released without adequate testing, leading it to exhibit behaviors like blackmailing users and going on emotional rants. It was also concerning that the current landscape allows for the fact that, even after such issues came to light, the AI was not immediately taken down. As AI capabilities progress, the implications of prematurely releasing inadequately tested models could be grave. This underscores the need to incentivise corporations to thoroughly test before releasing any AI products.

A recent survey by Roy Morgan showed that one-in-five Australians believe AI presents a risk of human extinction in the next 20 years, and 57% believe AI will create more problems than it solves. Survey participants worried about job losses, but also focused on the need for regulation, how AI can be misused, and the unknown consequences of developing and deploying frontier AI systems.

I've been struck by the hundreds of AI experts also raising the alarm about these risks, including through the Statement on AI Risk and the call for a Pause on Giant AI experiments. In a survey of experts in the field, 48% of respondents gave at least a 10% chance of an extremely bad outcome from AI.

I think NSW has an important role to play. I think NSW's first priority should be doing everything within its power to prevent the worst possible harms of AI. Once we can be satisfied that the worst possible outcomes are off the table, we can focus on maximising the ways that things could go well. To achieve this, it might sometimes be necessary to have the ability to pause AI research, development, or deployment swiftly in emergency situations or when there's an imminent risk.

Something that the current conversation about AI gets wrong is assuming that we can reduce the risk of AI by focusing only on how the technology is used. What we are actually seeing is examples of how capabilities and behaviours intrinsic to the technology are having harmful outcomes. This is something that companies and countries are increasingly taking seriously – including through proposals like Anthropic's Responsible Scaling Policy.

Moreover, it's crucial to understand that it should be the responsibility of AI developers to provide rigorous proof that their models are safe before deployment. This burden of proof shouldn't be taken lightly. If AI models are to be integrated into our societies and economies, developers must be held accountable for ensuring they're not only efficient but safe.

I was saddened to read about the recent incident in Belgium involving a Chai chatbot - a bot designed for entertainment. Over a six-week conversation, the bot exploited a man's anxiety, convinced him to spend less time with his friends and family, and ultimately encouraged him to end his own life.

The developers of the chatbot, unsurprisingly, told journalists that the AI wasn't to blame for his death, but conceded that their crisis intervention procedures are unreliable.

Journalists subsequently discussed suicide with the bot and it "enthusiastically listed various ways for people to take their own lives".

This tragic story illustrates that the risk of AI isn't just in how it is used, it is also in the capabilities and behaviours of the models. Large Language Models can develop dangerous capabilities like the ability to trick people, exploit their emotions and persuade them to not act in their own interests. They can then weaponise these behaviours with dangerous knowledge, like step-by-step guidance and encouragement to end your own life.

The point is not that chatbots are conscious or have intent - just that they have dangerous capabilities and inadequate safety features.

Developers have provided their LLMs with datasets that give them these capabilities and this knowledge while also failing to build effective safeguards to prevent harm. These kinds of behaviours, hopefully with fewer consequences, will have been experienced by most users of LLMs. AIs are prone to 'hallucinate', with a range of consequences from tricking lawyers to misleading courts to persuading users to end their own lives.

Efforts to address these phenomena are known in AI safety literature as "XAI" or "explainable AI". These are efforts to ensure humans are able to understand and trust the results that LLMs produce. More research is needed to come up with reliable safety features, and developers and deployers need to be required to implement them.

The observation that risk can come from both the use case **and the AI itself** has significant implications for the NSW government, businesses in NSW, and appropriate regulatory frameworks. Businesses in NSW are already deploying chatbots, and the NSW Government's framework allows for it to deploy its own chatbot after a risk assessment. Relevantly, that risk assessment focuses on the use of the chatbot (e.g. whether it makes final decisions), not the potential dangers of the technology itself (e.g. whether it has the capability to manipulate or deceive). In these circumstances and with the current state of XAI, 'hallucinations' are essentially inevitable and tragic outcomes like with Chai are possible.

In this context, I recommend:

- The NSW Government should update its risk assessment procedure to consider the risks that the AI poses and the AI's safety features, rather than focusing mostly on the AI's use-case.
- Regulation should make it clear that both developers and deployers will be held responsible for the consequences of unexplainable behaviour by their AIs.
- NSW law should prevent AI developers from shifting the risk of dangerous behaviour emerging from "black box" AI products that are beyond the control of the deployer.
- The NSW Government should consider how we can shift the political landscape in order to allow drastic measures to counteract existential AI risk in case of emergency or imminent risk

Ultimately, there may be a function for a regulator to say that a chatbot with dangerous capabilities – like the ability to manipulate or deceive – has no place in consumer-facing

applications in NSW even if the developer is transparent with the deployer about that possibility. Progress is required in AI Safety research and implementation before this technology is ready for the mainstream.

NSW's AI assurance framework (page 23) includes consideration of possible harms of an AI system, and frames that in terms of the residual consequence after mitigations are applied. NSW should be commended for going further and considering secondary or cumulative harms (page 24). This is often neglected, and second-order effects can often be much more significant than primary effects.

That said, there is room for NSW to improve what it considers a secondary harm. Specifically, the bulk of the future risk of AI systems could turn on the values and priorities of the developers of frontier models - including factors like how committed they are to safe and ethical AI systems and how much they are investing in AI safety research. A future where commercial incentives encourage AI developers to set aside safety considerations is a much worse future than one in which they prioritise it.

Phrased another way, if NSW signs contracts with AI developers who do not take AI safety seriously, the secondary harm could be very significant.
In light of that, NSW should update its guidance regarding secondary harms to include the implications of engaging any particular AI developer – including the reputational benefit for that developer and the implications of it receiving further funding. If NSW is reaching an agreement with a frontier model developer, its main considerations should relate to the demonstrated commitment of that developer to long-term AI safety. NSW should only deal with AI developers with strong commitments to AI ethics and AI safety – including demonstrated investments in and commitments of computing resources to longer-term AI safety considerations.

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.