

Submission in Response to the Artificial Intelligence: Governance and Leadership White Paper

To Whom It May Concern,

My name is [REDACTED] I am an Australian postgraduate student, currently conducting research on artificial intelligence at the Australian National University.

It's wonderfully encouraging to see that The Australian Human Rights Commission and the World Economic Forum are giving serious consideration to the opportunities and risks which AI poses for our nation. Thank you for the work you are doing in this crucial area, and for inviting submissions from the public.

Due to my specific domain knowledge, this submission is focused on the following consultation questions:

- 1 (*"What should be the main goals of government regulation in the area of artificial intelligence?"*)
- 6a (*"If Australia had a Responsible Innovation Organisation, what should be its overarching vision and core aims?"*)
- 6e (*"If Australia had a Responsible Innovation Organisation, how should it interact with other bodies with similar responsibilities?"*).

The long-term challenge posed by AI

The increasing capabilities of AI systems are likely to present risks and opportunities for Australia in both the near and far term. It can be tempting to focus exclusively on the short-term ramifications of improving AI technology, since they are more predictable and tractable than the long-term impacts. However, despite this, it is still in our nation's best interest to ensure we give sufficient consideration to less immediate-term concerns as well.

Most current AI systems are 'narrow' applications – specifically designed to tackle a well-specified problem in one domain, such as playing a particular game or classifying images. Such approaches cannot adapt to new or broader challenges without significant redesign. While the system may be far superior to human performance in one domain, it is not superior in other domains. However, a long-held goal in the field has been the development of artificial intelligence that can learn and adapt to a very broad range of challenges while operating in a wide range of environments. This kind of system is often referred to as Artificial General Intelligence [AGI].

Recent progress in the field of AGI has been encouraging: for one example, consider AlphaZero. AlphaZero is an AI system by London based AI research lab Google- DeepMind in 2018. AlphaZero was able to learn to outperform both human experts and game-specific algorithms in Go, Shogi and chess without having been specifically designed for any one of these games ([Silver et. al. 2017](#)).

This system was provided with no domain knowledge other than the rules of the game in question, and achieved these performance levels after several hours of playing only against itself. There is, of course, a huge gulf between an algorithm capable of learning multiple board games, and a system that approaches the level of general problem-solving ability of a human. However, there are likely to be further advances in research on systems that 'learn to learn' without being hand-crafted for a particular challenge, and many practical and scientific applications for such flexible, adaptable systems.

While it may seem unlikely to laypeople, many experts in the field believe there is a significant possibility that meaningful advances will be made in artificial general intelligence [AGI] within the next 40 years ([Grace et al. 2017](#)).

If artificial general intelligence is developed in the coming few decades, this will have a transformative impact the lives of all Australians (as well as the world at large).

The arguments for why we should expect a successful AGI project to have a dramatic impact on the world, and for why this impact isn't necessarily guaranteed to be positive, are quite involved. For in-depth discussion of this I would recommend referring to any of the following resources:

- (Book), [*Superintelligence: Paths, Dangers, Strategies*](#), Nick Bostrom, 2014, Oxford University Press;
- (Online Article), [*Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity*](#), Holden Karnofsky, 2016, for Open Philanthropy Project;
- (Online Article), [*Benefits & Risks of Artificial Inetlligence*](#), Max Tegmark, for Future of Life Institute; and
- (Online Article), [*Why AI Safety?*](#), The Machine Intelligence Research Institute.

An *extremely condensed* argument for why functioning AGI technology would be enormously impactful goes as follows.

[Most experts agree](#) that there is a moderate likelihood of developing goal-directed systems which outperform humans at a majority of tasks in the coming decades. Once we do build such systems, whatever goals these systems are given will be pursued extremely competently.

If the goals of these machines are aligned with the things we value, then the consequences will be extremely good. A superintelligent system would allow us to, for instance, accelerate scientific discovery, develop more efficient renewable energy technology, cure more diseases, run businesses more efficiently, and predict the consequences of public more reliably,

along with a whole host of other applications of which we are as-of-yet unaware. However, if the goals of these hyper-competent systems aren't aligned with our values, then these AI systems will seek to reshape the world according to these misaligned goals. Once such systems are released into the world, this cannot be reversed. If we cannot even prevent a [YouTube video](#) from proliferating on the internet, we have no hope of removing a super-competent adversary which is actively trying to thwart our efforts. The problem of how to align the goals of AI systems with our own is presently an [open unsolved problem](#) in the scientific literature.

How we ought to respond to these challenges

Nearly all researchers seriously working on artificial general intelligence with whom I have spoken agree that the challenges and risks posed narrow AI systems in the near-term (next few years), and the challenges posed by advanced general AI in coming decades are distinct problems which should be addressed separately using very different strategies.

Protecting Australian citizens from some of the near-term narrow-AI issues we're already seeing, or expect to see within a couple of years - such as machine learning systems exhibiting racial bias, firms with large data assets gaining monopoly power, or rapid disruptions to the Australian labor market - is crucially important for the health of our nation. It is commendable that these issues were identified in the white paper.

Although, in my view, many of these problems are deviously difficult, they can in principle be addressed domestically through new Australian laws, regulations, and government services.

Unfortunately, however, when it comes to mitigating the risks and capturing the benefits posed by very advanced AI in the less immediate term,

Australian domestic interventions alone might not be sufficient to protect the Australian people from the relevant risks. If a Responsible Innovation Organisation in Australia were to make a substantial difference in this area, the best way to have an impact likely involves Australia positioning itself as a leader and role-model in international collaboration efforts. The reason for this is that, even though our nation is exceptionally prosperous, intelligent, and vibrant, due to our relatively low population we accommodate only a tiny proportion of the world's top AI researchers. Our country's relatively small size also means we would find it very difficult to match the level of investment in AI research and development that we're seeing in countries such as USA, China and the United Kingdom. In 2015 the combined R&D budget of US AI-intensive firms was about [587.7B USD](#). This is over 40% of [Australia's total GDP in the same year](#). Since 2015, AI investment has been continued to sharply increase across the board. Since we can't possibly hope to match this level of investment, it's more likely than not that any particularly important discovery in the field of Artificial General Intelligence will come from a team working outside Australia.

This fact presents us with a challenge, since due to this emerging technology's potential to have globally transformative impacts, every serious artificial general intelligence project located anywhere on earth still poses risks to Australians, regardless of whether or not the research laboratory happens to reside on our continent.

In this way, Australia's relationship with the problems posed by advances in artificial general intelligence is somewhat analogous to her relationship to global problems with chlorofluorocarbon (CFC) emissions, or the proliferation of nuclear weapons. The best way for us to have an impact is to be actively participating in and promoting international collaborative efforts. Since the actions of every country have safety implications for citizens of all other countries, it is squarely in our own national interest to promote a healthy, cooperative, safety-conscious international community.

For this reason, if Australia had a Responsible Innovation Organisation, its overarching vision should include the goal of improving the level of international collaboration and coordination between governments around the world in the area of managing the potentially global risks posed by emerging technologies such as AGI. Correspondingly, A Responsible Innovation Organisation should also be focused promoting international efforts to ensure that the immense benefits offered by emerging technologies such as AGI will be justly distributed across all nations, including Australia.

An international strategic landscape which Australia should be aiming to avoid, is one where different teams competing to develop AGI are incentivized to adopt a policy of racing to deploy an AGI system as quickly as possible, before the safety of the system has been thoroughly established. This kind of arms race dynamic is tragic because, with better international cooperation and coordination, all stakeholders in this situation could be made better off if everyone was to show more restraint.

If an AI arms race intensifies between private firms (e.g., between Baidu and Google Deepmind), or between nation-states (e.g., China and the USA), where competing stakeholders act as if they are in a “winner take all” situation, this will lead to powerful incentives for competing teams to cut corners on safety and deploy potentially dangerous AGI systems before their risks have been properly explored.

As an example of the kind of collaborative attitude which Australia might foster, please refer to the charter of OpenAI. OpenAI are a widely respected AI laboratory in San-Francisco which is focused on ensuring AI has a beneficial impact on the world. In the last several years, OpenAI has been responsible for some of the most important results in the field of advanced AI development, such as their celebrated [GPT-2 language model](#). [The OpenAI charter states:](#)

"We are committed to doing the research required to make AGI safe, and to driving the broad adoption of such research across the AI community. We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions. Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project. We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions. Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project. We will work out specifics in case-by-case agreements, but a typical triggering condition might be "a better-than-even chance of success in the next two years."

Since the potential risks and benefits of this technology are immense for our nation, and for all others - and since the potential ramifications of this research will be unconstrained by national borders - Australia should become a leader in promoting safety-conscious international agreements aimed at avoiding dangerous arms-races, thus improving the chances of Australia eventually sharing in the monumental benefits which advanced general artificial intelligence technology might someday deliver.

Thank you for your attention, if wish, please feel free to contact me at

[REDACTED]

[REDACTED]

[REDACTED]