To the Select Committee on Adopting Artificial Intelligence,

My name is Mitchell Laughlin, a 24-year-old graduate of the University of Queensland, currently living in the electoral division of Grayndler, and employed at the Australian Department of the Treasury.

This letter represents my views as an individual and does not reflect the views of my employer. However, I am thankful for the support I received from The Good Ancestors Project and Effective Altruism Sydney in preparing this submission.

Public Concern on AI Safety

Research by Ready Research and The University of Queensland shows that most Australians are worried about AI safety. A common concern is that AI systems may not be safe, trustworthy, or aligned with human values. The majority of respondents felt that the Australian Government should focus on preventing dangerous and catastrophic outcomes from AI, require mandatory safety audits of new AI models before their release, and hold AI companies accountable for any harm they cause.

In my personal experience, everyday Australians echo these concerns about AI safety, with worries ranging from increased cyber-attacks and the creation of novel biological weapons to labs losing control of advanced autonomous AI systems not aligned with human values. The Center for AI Safety's Statement on AI Risk, which is endorsed by world-leading experts like Turing Prize winners Geoffrey Hinton and Yoshua Bengio, underlines these concerns.

While I commend the Government's focus on driving AI adoption through initiatives like CSIRO's National AI Centre, I believe it is equally important to address bigger risks. The scope of these risks is beyond the capacity of individuals and requires the concerted action of governments. I urge this Committee to recommend the Government prioritise AI Safety issues alongside its current focus on AI adoption and addressing immediate problems posed by AI.

Creation of an Australian AI Safety Institute

Australia should establish a National AI Safety Institute, similar to those already existent in countries such as the US, UK, Canada, and Japan.

Drawing inspiration from the UK AI Safety Institute, our Australian equivalent would focus on three core priorities:

1. **Evaluating advanced AI systems.** This involves reviewing the characteristics and capabilities of new AI systems, understanding the adequacy of their safeguards, and considering any implications or impacts they might have. The process of 'red-teaming', or having trusted experts attempt to bypass safeguards or identify potential dangers, would be integral to this evaluation.

2. **Drive foundational AI safety research.** In order to protect public interest, research on understanding AI systems and ensuring their safety needs to keep pace with the swift advancement of AI capabilities driven largely by private sector investment. The Institute could coordinate research agendas to ensure safety is given due consideration and that capabilities do not supersede controls.

3. **Engage in national and international partnerships on AI Safety.** By participating in partnership agreements similar to those announced by international labs, Australia could contribute to exchanging methodologies, conducting personnel exchanges, assisting in standards development, and collaborating in joint testing. Information-sharing with other national and international actors, including policymakers, private companies, academia, civil society, and the broader public, would also be facilitated by the Institute.

By establishing an AI Safety Institute, we would not only be preparing for any future regulatory regime but also developing the technical capability and capacity to administer that legislation. This would provide Australia with more information and options when it comes to implementing regulations.

I firmly believe a top priority of the Australian Government should be preventing dangerous or catastrophic outcomes from AI.

Thank you for the opportunity to contribute to this important discussion.

Regards,
Mitchell Laughlin