Opening statement from Mr Sadler before the Senate Committee on Adopting AI on 16/8/2024:

Good Ancestors wants to help us all be good ancestors to future generations by making forward-looking, evidence-based, and practical policy recommendations.

As part of that goal, I coordinated the submission from Australians for AI Safety. Soroush was one of its signatories.

When we talk about safety, we mean physical harm to people or infrastructure, especially at a large scale. The same meaning of “safety” as in “aviation safety”.

These are the kinds of issues that the CEOs of all the top AI companies; and key academics; raised last year in widely publicised open letters.

Submissions to this inquiry, the public, and experts overwhelmingly agree that safety needs to be a priority.

I want to focus on what those safety concerns are and how we can practically address them.

To make good AI policy, we need to think about the models of tomorrow.

In the Good Ancestors Submission, we present evidence that AI capabilities have grown rapidly and unpredictably. That capability growth is likely to continue or accelerate.

And we should anticipate that some of those new capabilities will be dangerous.

This could be because they are misused, or… in the future… we could lose control of highly capable AI systems.

Biosecurity is perhaps the most useful example to start understanding misuse risks.

In 2017, a leading laboratory manufactured an extinct relative of the smallpox virus using DNA ordered online.

Today’s AIs are on the cusp of being able to help thousands of people do what was previously only possible in leading labs.

In March 2022, a paper was published in Nature Machine Intelligence detailing how an AI intended to find new drugs instead designed 40,000 novel and lethal molecules in less than six hours.

Similarly, a 2023 study showed that students were able to use ChatGPT to:

1. Suggest potential pandemic pathogens
2. Explain how they can be made from DNA ordered online, and

3. Supply the names of DNA synthesis companies unlikely to screen online orders to ensure they don't include dangerous DNA sequences
4. The study also showed that ChatGPT's safeguards failed to prevent it from providing this dangerous assistance.

It's realistic that AI could help people build bioweapons during the next term of government. The US has taken action to address this safety risk, including through an executive order in October 2023.

I've raised this issue with several Departments but have not seen any evidence of ownership of this risk in the Australian system; or intent to follow the US in implementing targeted safeguards.

Similar misuse concerns arise from other capabilities, such as cyber-offence. While GPT3.5 had limited cyber offensive capability, a series of papers published earlier this year showed that GPT-4 was able to autonomously hack websites and exploit 87% of newly discovered vulnerabilities in real-world systems.

If developers create future generations of AI systems with advanced cyber offensive capabilities and inadequate safeguards, it would dramatically change the cyber security landscape.

But there are practical things we can do to address safety concerns.

Establishing an Australian AI Safety Institute, based on the UK model, that focuses on frontier AI safety is the first step.

An AI Safety Institute would help tackle uncertainty by contributing to the evaluations necessary to understand next-generation systems and brief Australian policy-makers accordingly.

Finally, we have useful overseas models that can help us regulate AI in Australia.

- Canada and the EU have useful definitions for "high-risk AI" and "general-purpose AI with systematic risks" that will let us target regulation only at risky models and avoid harmful overregulation.
- Californian Bill SB-1047 has specific proposals that we could adapt, like:
    - Requiring the most advanced AI models to undergo safety testing and be available to third parties for evaluation.
    - Requiring developers and deployers to be able to shut down AI models in certain circumstances, and
    - Mandating safety incident reporting, like in the aviation industry.

We have to get ahead of these problems, rather than taking a "wait and see" approach that puts an unacceptable risk onto future generations.