AI Safety Australia & New Zealand

AI Safety Australia & New Zealand is a non-profit organisation dedicated to addressing the challenges and potential risks associated with the development of advanced artificial intelligence systems. Our mission is to grow and support a large, ambitious, and influential local community focused on preventing the most harmful impacts of AI.

Key activities of our organisation include:

1. **Community Building**: We foster a network with over 500 members across Australia and New Zealand through online forums and regular in-person meetups in major cities.

2. **Education and Career Development**: We provide resources, career coaching, and organise the region's largest annual conference on AI safety careers.

3. **Policy Engagement**: We actively participate in policy discussions and consultations to promote responsible AI development and governance.

4. **Talent Pipeline**: We connect high-potential individuals with educational, job, networking, and funding opportunities in the field of AI safety across Australia and New Zealand.

Our organisation brings together a diverse group of professionals, researchers, and concerned citizens who are committed to ensuring that the development of powerful AI systems aligns with human values and safety considerations. We believe that establishing mandatory guardrails for AI is a critical step in achieving this goal.

The following recommendations reflect the collective expertise and concerns of our community regarding the proposed AI regulations.

Recommendations

- Trust in AI systems is low[1] because the Government is increasingly falling behind the rapid advance in the risks of AI models.[2] Many of our members were concerned about the risks of advanced AI systems well before the open letters from experts in March and

[1] M Noetel, A Saeri and J Graham, '80% of Australians Think AI Risk is a Global Priority – Government Needs to Step Up' (Article, 11 March 2024). https://www.uq.edu.au/research/article/2024/03/80-australians-think-ai-risk-global-priority-government-needs-step.
[2] *OpenAI*, 'OpenAI-01 System Card' (Web Page, 2023) https://openai.com/index/openai-o1-system-card/.

May 2023 and the US Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence in October 2023. It's now October 2024, and Australia has not only failed to implement the basic risk mitigations in the US Executive Order, but we've fallen even further behind. **We recommend that Australia take immediate action *before the next election* on pressing risks – like the possibility that advanced AI systems could be used to make bioweapons or cyberweapons.** This action can use existing powers and does not have to wait for a new regulatory scheme (Question 16).

- Our experience is that there is a large and skilled Australian community concerned about the safety of advanced AI systems,[3] but the pipeline to get the talent working on the problem is lacking. Many of our members have the skills and desire to work on reducing catastrophic risks from AI systems, but Australia's current technical and regulatory landscape offers few jobs of this kind. Many of our members have left the country to work on these problems in the US and UK. **We recommend that the Government design safety regulations that leverage Australia's AI safety workforce.** This could include (Questions 15 and 16):
    - **Ensuring appropriate focus on possible catastrophic risks across the regulatory environment**
    - **Giving a future Australian AI regulator specific powers relating to the safe development of the most advanced AI systems, and**
    - **Ensuring that Australia has an AI Safety Institute to support work across the public sector.**

- Large offshore developers control most AI risks. Front of mind in any AI regulation should be *what* obligations Australia can impose on those developers and *how* they can be effectively enforced. The current approach to defining high-risk AI combines powerful AI models being developed by these companies with far less risky AI systems being deployed in Australia. The current approach also applies a single set of guardrails across this range of systems and models. We risk losing track of what matters – a small number of powerful AI models – by hiding them among much more numerous and much less risky AI systems (Questions 7, 10, 11 and 13) **We recommend:**
    - **The riskiest AI systems be identified by their own specific definition, perhaps drawing on FLOP thresholds like the EU AI Act or monetary thresholds like SB1047.**

[3] Ben Abbott, *Australia Could Have 200,000 AI Tech Workers by 2030* (31 July 2024) *TechRepublic* https://www.techrepublic.com/article/tech-council-australia-ai-job-predictions.

    - **The riskiest AI systems be subject to their own specific guardrails, perhaps drawing on SB1047 and the recommendations of GovAI.**[4]
    - **An Australian AI Act explicitly direct the new AI regulator to "triage" the challenges of AI – ensuring that we appropriately prioritise risks that could have catastrophic consequences before moving on to smaller-scale harms.**
        - Empowering established regulators to focus on the deployment of AI systems within their spheres of responsibility may be the first step in allowing a new AI regulator to focus on larger-scale risks.

- An Australian AI Act needs to be drafted with the future in mind. Many of our members think that artificial general intelligence is a realistic possibility this decade. The safeguards we have today, in Australia and globally, are manifestly inadequate for this possibility. Any legislation that results from this consultation is very likely to still be in force when AGI is developed. **We recommend that the Department of Industry perform scenario-based stress testing of any legislation to ensure that it is sufficiently flexible and sufficiently powerful to navigate step-changes in AI capability, like the development of AGI** (Question 15).

Yanni Kyriacos
Co-Founder & Director
AI Safety - Australia & New Zealand

[4] J Schuett, N Dreksler, M Anderljung, D McCaffary, L Heim, E Bluemke and B Garfinkel, *Towards Best Practices in AGI Safety and Governance: A Survey of Expert Opinion* (Centre for the Governance of AI, May 2023).