Australians
For AI Safety

Dear Minister,

We write to call for the Australian Government to take the risks of AI seriously.

The economic potential of advanced AI systems will only be realised if we make them ethical and safe. Ethical challenges from today's systems are already causing serious harms, and as systems become more powerful the risks of misuse, accident, and catastrophe become more acute.

The safe use of AI requires considering that, in the future, AI could represent a catastrophic or existential risk that could jeopardise all of humanity.[1] While we have to be frank about the uncertainty, many experts have raised the alarm, and governments must listen.[2] Mitigating catastrophic risk should never be left to chance.

As part of a due-diligence-based approach, the Australian Government should:

Recognise the risk. The Australian Government's AI strategy must recognise that catastrophic and existential consequences are possible. The AI strategy should highlight that mature risk management processes treat uncertainty as a cause for concern, not comfort.

Take a portfolio approach to mitigating risk. The Australian Government's risk-based approach to AI should be holistic, including preparing for problems that may only arise when AI systems are more advanced[3]. Mitigations need to be in place for the risks we are experiencing today, like targeted harassment, dual-use technologies, deepfakes, and other forms of misinformation and disinformation.[4] Risks with uncertain likelihood, but catastrophic consequences, must also be mitigated within the portfolio.[5]

Work with the global community. The rest of the world is moving quickly, and we should contribute. Australia has led on the risks of nuclear and biological weapons – we can and should lead on mitigating similar risks from AI.[6] Specifically, our contributions to international agreements and standards development should be mindful of managing longer-term and catastrophic risks while also addressing ethical concerns and ensuring economic benefits.

Support research into AI safety. In addition to policy leadership, we need technical solutions that greatly improve AI interpretability and alignment to human values. We must not leave the risks of AI to chance – or private interests and foreign companies. This requires governments to support research into AI safety, and urgently train the AI safety auditors that industry will soon need.[7] Technical problems take time to solve, so we need to start now.

A powerful first step from this consultation would be the creation of an AI Commission, or similar body, tasked with ensuring Australia takes a world-leading approach to understanding and mitigating the full range of risks posed by AI.

An AI Commission should approach its work on AI safety in a similar way to how we approach aviation safety. Government doesn't need all the technical answers – but it does need to set the expectation of a culture of safety and transparency, create an independent investigator and a strong regulator, back them with a robust legal regime, and connect them with a global peak body that our Government helps to shape. This is how the Australian Transport Safety Bureau, the Civil Aviation Safety Authority and the International Civil Aviation Organization give Australians the confidence to fly. An AI Commission must ensure AI has the same kind of governance so that Australians can give it the same kind of trust.

An AI Commission should be set up urgently to ensure the law is clear about who is liable for harms caused by AI. This should include joint culpability between AI labs and AI providers, as well as anyone who uses AI to cause harm. We wouldn't allow aircraft manufacturers to sell planes in Australia without knowing the product is safe, and we wouldn't excuse a business for being ignorant about the potential harms of its products, so the law should similarly ensure adequate legal responsibility for the harms of AI.[8]

Importantly, this globally coordinated regulatory approach to aviation hasn't stifled innovation. Indeed, certainty and structure of this kind helps new participants by providing them a framework through which to participate.

The undersigned individuals and organisations represent a cross-section of Australian AI expertise. We call for Australia to listen to experts raising the alarm about the catastrophic risks advanced AI could bring, and to do the due diligence necessary to ensure humanity follows a safe path.

Australians For AI Safety

Soroush Pour
CEO
Harmony Intelligence

Engineer & Technology Entrepreneur

JJ Hepburn
CEO
AI Safety Support

Chris Leong and Yanni Kyriacos
Convenors
AI Safety Australia and New Zealand

Dr Daniel Murfet
Lecturer in Mathematics
University of Melbourne

Hunter Jay
CEO
Ripe Robotics

Michael A Osborne
Professor of Machine Learning
University of Oxford

Simon Goldstein
Associate Professor
Dianoia Institute of Philosophy,
Australian Catholic University

Research Fellow
Center for AI Safety

Richard Dazeley
Professor of Artificial Intelligence & Machine Learning
Deakin University

Leader
The Machine Intelligence Lab

Senior Member
The Future of Life Institute AI Existential Safety Community

Co-founder
The Australian Responsible Autonomous Agents Collective (ARAAC.au)

Michael Dello-Iacovo
Strategy Lead and Researcher
Sentience Institute

Dan Braun
Lead Engineer
Apollo Research

Matthew Farrugia-Roberts
AI Safety Researcher

Dane Sherburn
AI Safety Researcher

Joseph Bloom
AI Safety Researcher

Peter Vamplew
Professor of Information Technology
Federation University

Senior Member
The Future of Life AI Existential Safety Community

Hadassah Harland
PhD Candidate, Artificial Intelligence and Human Alignment
Deakin University

Member
Australian Responsible Autonomous Agents Collective (ARAAC.au)

Harriet Farlow
CEO
Mileva Security Labs

PhD Candidate, Adversarial Machine Learning
UNSW Canberra

Michael Aird
Senior Research Manager, AI Governance and Strategy
Rethink Priorities

Toby Ord
Senior Research Fellow
Future of Humanity Institute
Oxford University

Australians
For AI Safety

Endnotes

[1] Hendrycks, Mazeika, & Woodside. "An Overview of Catastrophic AI Risks." June 2023. arXiv preprint arXiv:2306.12001.

[2a] "Statement on AI Risk." Centre for AI Safety. May, 2023. safe.ai/statement-on-ai-risk

[2b] "Pause Giant AI Experiments: An Open Letter." Future of Life Institute. March 2023. futureoflife.org/open-letter/pause-giant-ai-experiments

[2c] "The Godfather of A.I.' Leaves Google and Warns of Danger Ahead." New York Times, 1 May 2023. nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

[2d] "AI 'godfather' Yoshua Bengio feels 'lost' over life's work". BBC. May 2023. https://www.bbc.com/news/technology-65760449

[2e] "In a survey of AI researchers carried out in 2022, 48% thought there was at least a 10% chance that AI's impact would be 'extremely bad (eg, human extinction)'". The Economist. April 2023. https://www.economist.com/leaders/2023/04/20/how-to-worry-wisely-about-artificial-intelligence

[3] Whittlestone & Clark. "Why and How Governments Should Monitor AI Development." August 2021. arXiv preprint arXiv:2108.12427

[4a] Soice et al. "Can large language models democratize access to dual-use biotechnology?" June 2023. arXiv preprint arXiv:2306.03809

[4b] Eshoo. "Eshoo Urges NSA & OSTP to Address Biosecurity Risks Caused by AI." October 2022. eshoo.house.gov/media/press-releases/eshoo-urges-nsa-ostp-address-biosecurity-risks-caused-ai

[5] Shevlane et al. "Model evaluation for extreme risks." May 2023. arXiv preprint arXiv:2305.15324

[6] "Australia: A Leader in Global Statistics and Rankings." Australian Government Department of Foreign Affairs and Trade. globalaustralia.gov.au/why-australia/statistics-and-rankings

[7a] Falco et al. "Governing AI safety through independent audits." Nature Machine Intelligence 3, no. 7 (2021): 566-571. doi: 10.1038/s42256-021-00370-7

[7b] Avin et al. "Filling gaps in trustworthy development of AI." Science 374, no. 6573 (2021): 1327-1329. doi: 10.1126/science.abi7176

[8] "Strong and appropriate regulation of advanced AI to protect humanity." Campaign for AI Safety. April 2023. campaignforaisafety.org