Friday July 28th, 2023

Dear Minister,

I am an academic researcher aiming to identify, understand, and reduce risks of harm to society from advanced intelligent systems. I write in response to the Government's recent discussion paper titled "Safe and Responsible AI in Australia".[1]

I am pleased that the Government is taking a serious approach to addressing the many harms already arising from artificial intelligence (AI) systems. These harms are real. Their mitigation is urgent and important work. I applaud the Government's movements in this direction.

I noticed that the paper leaves out of scope a discussion of certain broader impacts of AI, including on the labour market, intellectual property, national security, and the military. I look forward to further discussion of these urgent and important issues in the near future.

I remain deeply concerned that the Government's model of specific risks from AI systems is, so far, too narrowly focused on too small a class of risks to qualify as truly *responsible*. I implore the Government to broaden its definitions and taxonomies of risk in anticipation of additional[2] pathways to future harm, including:

1. pathways to harm that are not currently prevalent but are likely to become prevalent soon, such as harms from interactions with adversarial autonomous AI agents;[3]
2. pathways to harm that are currently considered uncertain or speculative but may have catastrophic scale, such as existential risks from uncontrollable AI systems;[4] and
3. pathways to harm that are currently unknown to us.

These pathways to harm are less obvious and well-understood compared to the clear and immediate risks identified in the discussion paper. Experts do not universally agree on whether these uncertain harms will eventuate.[5] However, *this does not mean that they should be dismissed.*

On the contrary, dismissing these risks because we don't understand or agree upon them would be highly irresponsible, since there is some chance that they will arise rapidly and unpredictably; and there is some chance that they will lead to devastating harm. When facing uncertainty about potential future harms, the responsible thing to do is to investigate the potential harms in proportion to their plausibility and their scale.

In the present moment, we are witnessing the weaving of global narratives about the place of AI in our future world, and the path society should take to get there. The Government has taken the laudable step of acknowledging the crucial importance of responsible management, and opening its eyes to the real harms caused by today's AI systems. I implore the Government to take this opportunity to lead the world in acknowledging and responsibly navigating the *full scale* of risks from AI systems. To do so requires *at least* the following:

1. funding interdisciplinary research identifying potential harms from future AI systems, reducing uncertainty about these harms, or reducing their chance of being actualised;
2. adopting an agile regulatory stance, addressing new risks as they come into focus; and
3. using Australia's diplomatic standing to spearhead global coordination on addressing international-level risks from AI systems.

Together, these steps constitute the start of a truly responsible approach to AI advancement.

Yours faithfully,

Matthew Farrugia-Roberts
Research Assistant in Human-Agent Interaction and Teaching Assistant in the Ethics of AI
School of Computing and Information Systems
The University of Melbourne

Notes and references.

[1]Department of Industry, Science and Resources (DISR), June 2023, "Safe and responsible AI in Australia: discussion paper".

[2]I must emphasise: I am *not* downplaying the real and immediate risks from present-day AI systems—such as those outlined in the discussion paper—in favour of future, uncertain, higher-scale risks. Rather, to safely realise the benefits of AI, we *must,* at a *minimum,* acknowledge and address *both* classes of risks.

[3]See, for example: Chan *et al.*, June 2023, "Harms from increasingly agentic algorithmic systems", in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.* DOI 10.1145/3593013.3594033.

[4]See, for example: Critch and Krueger, 2020, "AI Research Considerations for Human Existential Safety (ARCHES)," *preprint* arXiv:2006.04948. DOI 10.48550/arXiv.2006.04948.

[5]Over recent years a number of highly-credible AI experts have expressed concern about catastrophic and existential risk from AI, including 2018 Turing Award co-recipients and pioneers of modern AI, Geoffrey Hinton[6] and Yoshua Bengio,[7] and co-author of the standard introductory textbook on AI, Stuart J. Russell.[8,9] See also the recent "Statement on AI Risk" from the Center for AI Safety.[10] Other experts in AI and other fields have expressed scepticism or have outright dismissed the potential for catastrophic harms (including, notably, the third co-recipient of the 2018 Turing Award, Yann LeCun). However, to my point, this indicates uncertainty, not safety.

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

[7]July 2023, testimony of Professor Yoshua Bengio before the U.S. Senate Judiciary Subcommittee on Privacy, Technology, and the Law. https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-principles-for-regulation

[8]Russell, 2019, *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking.

[9]July 2023, testimony of Professor Stuart J. Russell before the U.S. Senate Judiciary Subcommittee on Privacy, Technology, and the Law. https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-principles-for-regulation

[10]Center for AI Safety, "Statement on AI Risk: AI experts and public figures express their concern about AI risk." https://www.safe.ai/statement-on-ai-risk. The full statement, signed by hundreds of AI experts and other public figures, is: "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."