This document contains a response to the *Safe and Responsible AI in Australia* discussion paper published in June 2023 by the Australian Government Department of Industry, Science and Resources. The response was written by:

- **Luke Thorburn**, an (Australian) PhD student in safe and trusted AI at King's College London, whose research focuses on the intersection of algorithmic recommender systems and societal conflict. He is a member of the GETTING-Plurality Research Network within the Edmund & Lily Safra Center for Ethics at Harvard University, and a research affiliate in the Machine Intelligence and Normative Theory Lab at ANU. He also co-authors the Understanding Recommenders project at the Center for Human-Compatible AI at UC Berkeley — which was cited in the recent US Supreme Court case *Gonzalez v Google* — and has worked with Ofcom on methods for evaluating recommender systems.
- **Thorin Bristow**, a (British) former Google DeepMind Scholar who works on AI governance, particularly in relation to the impacts of AI on socioeconomic inequality.
- **Liam Carroll**, an (Australian) mathematician and AI researcher whose work (on singular learning theory) has helped open a major research direction in technical AI safety called developmental interpretability.

Correspondence about this submission should be directed to Luke Thorburn at luke.thorburn@kcl.ac.uk.
We thank those in our academic networks who offered feedback drafts of this submission.

Definitions

1. Do you agree with the definitions in this discussion paper? If not, what definitions do you prefer and why?

We caution against the use of the terms "misinformation" and "disinformation". While in principle these terms are straightforward to define in terms of false information and the intent with which it is shared, to operationalise any of these definitions in legislation or code requires one to make contestable, labour-intensive decisions about what is and isn't true. This is impossible to do well at scale. A strong indication of the difficulties in assessing truth is that the main organ of the US intelligence community tasked with monitoring the quality of intelligence work — the Analytic Standards and Integrity (AIS) division of the Office of the Director of National Intelligence — does not include accuracy among the criteria it evaluates. Dr Barry Zulauf, former Chief of AIS, states:

> We in AIS have not evaluated products for accuracy for 6 or 7 years. In order to put judgements aside to test for later accuracy, they had to be clearly stated, falsifiable, and include a timeframe. THEN we had to devote personnel to doing the research through other reporting to assess accuracy, taking personnel away from the main line of work. AIS has steadily declined in personnel resources for the past 6 years, and has done NO such evaluation. Before, only a small proportion of the products we sample, which, in turn was a representative cross-section, not a statistically significant sample of all production, was evaluated for accuracy.[1]

[1] Barry Zulauf (June 2022). The quote is from private communication but Barry has given permission for it to be shared.

If the US intelligence community finds it difficult to evaluate the truth of their products, when they have strong incentives to be accurate, then it is implausible that the Australian government could evaluate accuracy at scale.

Recommendations

- Do not place government entities in a position where they are responsible for policing (and hence defining) “misinformation” or “disinformation”.

Potential gaps in approaches

2. What potential risks from AI are not covered by Australia’s existing regulatory approaches? Do you have suggestions for possible regulatory action to mitigate these risks?

The emergence of capable generative AI models has rapidly expanded the space of possible risks. It is not clear whether Australia’s existing regulatory approaches are sufficient to address many of these risks. For example, potential risks from AI include (among many others):

- Unintended harms from actors who are naïve of AI risks
- Adverse feedback loops between human users and algorithmic systems
- Production of novel pathogens
- Proliferation of fraud
- Concentration of power and financial windfalls from AI technology

In some of these cases there are existing regulations that “cover” the risk (e.g., fraud) but it is not clear whether existing enforcement mechanisms and judicial procedures are sufficiently well-equipped to be able to handle a rapid increase in the scale of harm (e.g. fraud enabled by emerging tools like FraudGPT and DarkBERT).

In other cases — such as adverse feedback loops between human users and algorithmic systems — it is not clear what existing regulation would be applicable. For example it is possible that a capable recommender system or large language model, in the course of optimising for a measure of engagement or user retention, may learn to systematically manipulate users’ preferences. It is not clear what existing regulation would prevent such outcomes.

Generally speaking, existing regulatory approaches in Australia also seem targeted at several *specific* risks, but are less well equipped to deal with *systemic* risks, such as concentrations of power that may result from AI monopolies, but also other systemic risks such as apathy towards the idea of “truth” as it becomes impossible to distinguish audio, video or photographic evidence that is real from that which is synthetic, or the (highly uncertain) risk of creating an uncontrollable autonomous agent whose interests are not aligned with our own. These kinds of systemic risks are also missing from the draft risk management approach described in Box 4 of the discussion paper.

Recommendations

- At minimum, ensure that users of AI systems know that they are interacting with an AI, and understand its limitations:
    - Require companies developing and deploying LLMs to post visible consumer-facing information on their interface so users are aware they are interacting with AI systems or generated content, with clear information about their particular system's hallucination prevalence and risk. Consider this analogous to FDA reporting requirements for food and drug allergens and side effects (to use a US example).
    - As part of the regulatory process, require firms to develop educational materials on the specific strengths and weaknesses of particular AI models as they are released (perhaps following prerelease audits and/or reporting) and develop materials on prompt-engineering best practices; these materials might be modelled on the educational materials that accompany the release of new pharmaceuticals.
- "War plan" scenarios in which existing regulation may appear to cover certain risks, but in practice not be effective as a mitigation measure due to increased pace or scale that is increasingly achievable to bad actors.
- Consider whether new regulations are necessary to protect people from risks that seem unique to AI systems, such as the risk of adverse human-machine feedback loops.
- Include *systemic* risks in any risk-based regulatory approach.

3. Are there any further non-regulatory initiatives the Australian Government could implement to support responsible AI practices in Australia? Please describe these and their benefits or impacts.

First, the Government could support public education about the limitations of AI systems. We frequently come across cases of well-meaning individuals deciding to implement some form of AI system — such as using language models to automatically grade postgraduate university applications — with simply no awareness that such systems may be biased or discriminatory. With greater awareness of the limitations of such technologies among individuals who are acting in good faith, considerable harm could be avoided.

Second, the Government could fund research into responsible AI practices, to help grow a body of expertise within Australia and limit a brain drain of many of the most talented AI researchers and ethicists moving overseas.

Recommendations

- Consider updating school and university curricula to include modules on the basics of AI.
- Set aside a percentage of federal AI research funding for research into ethical, legal and social implications.

4. Do you have suggestions on coordination of AI governance across government? Please outline the goals that any coordination mechanisms could achieve and how they could influence the development and uptake of AI in Australia.

We think one promising model is specialisation. Specifically, to create an in-government, public service consultancy akin to the structure of the Office of the Australian Government Actuary, or something analogous to the recently announced UK Civic AI Observatory (but within government). Such an organisation would be able to provide consistent advice across government departments, and have a concentration of relevant expertise that is much harder to achieve when departments are working in isolation. This model was recently advocated for by a number of leading Australian AI researchers and developers.

Recommendations

- Consider creating an in-government AI governance consultancy that can develop best practice, hire relevant experts, and coordinate AI governance efforts across other parts of government.

Responses suitable for Australia

5. Are there any governance measures being taken or considered by other countries (including any not discussed in this paper) that are relevant, adaptable and desirable for Australia?

A major category of governance measures not discussed in the paper is that of mechanisms for providing democratic oversight of AI systems and AI governance. Recent and ongoing advances in generative AI seem likely to cause enormous societal impacts, and it is right that society be consulted and involved in governing and steering these impacts. Australia has a strong foundation in democratic innovation, notably the work of the newDemocracy Foundation who have previously been involved in work providing democratic input to governance of tech platforms. Both local and state governments in Australia have previously convened citizen assemblies on contested topics.

Overseas, particularly in the US, there are increasingly large-scale experiments with democratic oversight methods for AI systems. These include the alignment assemblies currently being conducted by the Collective Intelligence Project, and the recently-awarded OpenAI grants for 10 groups to trial innovative methods for providing democratic input into AI systems.

Recommendations

- Convene, with expert help, participatory democratic and deliberative processes such as citizens assemblies, to assist in the development of AI regulation.

Target areas

8. In what circumstances are generic solutions to the risks of AI most valuable? And in what circumstances are technology-specific solutions better? Please provide some examples.

In general, technology-agnostic regulation tends to be more robust and relevant in the long term than regulation that is framed in terms of particular technologies. Thus, wherever possible, the focus should be

on regulating the *outcomes* we don't want (systematic bias, discrimination, lack of agency, physical harm, etc.), or on the *processes* of technological development that are considered ethically acceptable or unacceptable (e.g., specifying the due diligence that we expect those who develop or deploy AI systems to conduct in the process of their work). In some cases, such as recommender systems, it may be much easier to articulate ethically acceptable development processes than to dictate in advance the properties ethically acceptable algorithmic products. This is also true because large scale online algorithmic systems are often implicated in many different kinds of harm, and reducing some of these harms increases others. Thus, rather than banning harms which may not feasibly be prevented, it may be more fruitful to regulate the way in which different varieties of harm or risk are traded-off against one another.

That acknowledged, there may be some situations in which technology-specific approaches may be required. For example, it is possible that specifications of best practice designs of AI systems — with respect to ensuring that they are safe, trustworthy, or conform to legislation — may by necessity be specified in technology specific ways. At the present, we do not believe the government needs to specify such best practices, though it may become necessary as the capability of AI systems continues to grow.

Recommendations

- Wherever possible, write regulations in ways that are technology-agnostic.
- Write regulations which acknowledge that in large-scale online algorithmic systems (such as recommender systems or language model APIs), sometimes different varieties of risk or harm cannot be simultaneously reduced, but only balanced against one another.

11. What initiatives or government action can increase public trust in AI deployment to encourage more people to use AI?

Encouraging more people to use AI — with the motivation of expediting uptake of these technologies — should not be a government priority. The speed of transition will create significant challenges to governance and our ability to collectively adapt as things stand already, without additional pressures to accelerate. Generally speaking, the government should not arbitrarily obstruct or hamper adoption, but nor should they seek to actively accelerate it.

That said, the government may wish to take steps to improve access to these technologies among disadvantaged groups, or to direct resources towards applications of AI which are undervalued by the free market. Concretely, this may involve taking the following actions.

Recommendations

- Prioritise the use of AI in areas where workforce shortages cannot be otherwise mitigated.
- Invest in the long-term use of AI for augmenting existing jobs and worker productivity, especially for novice and low-skilled workers.
- Promote affordable access to core technologies (LLMs) through effective antitrust/competition law enforcement, and consider subsidising or providing publicly funded LLMs to those who would not otherwise be able to afford access.

Risk-based approaches

15. What do you see as the main benefits or limitations of a risk-based approach? How can any limitations be overcome?

The main benefit of a risk-based approach is that it is, at least in theory, minimally interventionist or bureaucratic, only imposing additional requirements when the level of risk requires it. The main limitation of this approach is that determination of the risk category into which a given use of AI falls will often be subjective or uncertain, and risk bands may be too coarse a categorisation to express the mitigations most appropriate in many contexts.

A version of a risk-based approach that relies less on predetermined risk bands and more flexibly allows for experimentation is described in a recent paper on *Frontier AI Regulation* and a corresponding blueprint for what such a process might look like in practice. For managing risks from new kinds of AI systems, or newly large AI systems, this kind of rigorous process might mitigate the biggest risks.

17. What elements should be in a risk-based approach for addressing potential AI risks? Do you support the elements presented in Attachment C?

Broadly, we support the elements in Attachment C, but acknowledge that they are in the current document quite vague.

There are reasonably mature efforts to articulate how many of these elements should be implemented. For example, in the context of reporting/accountability, please see work on data sheets, model cards, and reward reports (also see here).

19. How might a risk-based approach apply to general purpose AI systems, such as large language models (LLMs) or multimodal foundation models (MFMs)?

The risk framework may be applied not to specific technologies, but to specific uses or applications of a technology. For example, use of a language model to summarise search results is likely a lower risk application of the technology than use of a language model to build an autonomous goal-directed system similar to design of AutoGPT or BabyAGI.

20. Should a risk-based approach for responsible AI be a voluntary or self-regulation tool or be mandated through regulation? And should it apply to:

a. public or private organisations or both?
b. developers or deployers or both?

Above a given threshold of risk, it should be mandated through regulation. It should apply to both public and private organisations, to deployers, and for some kinds of risks, to developers.

To clarify that last point, the regulation should apply to developers only insofar as there is risk inherent in the development process. For example, developers of a biased facial recognition algorithm should, assuming they have honestly communicated its shortcomings, not be held responsible for the decision to deploy it, and any harms that result (in this scenario, it should be the deployers who are at fault). However, some risks — such as the risk of synthesising contagious lethal pathogens or the risk of out-of-control

autonomous systems — may be inherent to the development process of some kinds of AI systems, and in situations where this is the case, developers should also be subject to risk-based regulation.