Centre for
Digital Wellbeing

Safe AI in Australia: Proposed Framework for Responsible Use

Submission

28 July 2023

Table of Contents

Australian Government
Department of Industry, Science and Resources
Technology Strategy Branch
DigitalEconomy@industry.gov.au

28 July 2023

Submission on safe AI regulation and governance

Introduction

The Centre for Digital Wellbeing (CDW) welcomes the opportunity to provide a submission to the Department of Industry, Science and Resources regarding actions that can be taken on AI regulation and governance, responding to the 'Safe and responsible AI in Australia Discussion Paper' (June 2023).

The Centre is a policy research and design body focusing on technology's impact on overall health and wellbeing, safety, and social cohesion in the Australian community. Our purpose is to facilitate research about the impacts of technology, formulate policy responses and develop initiatives to assist users to better engage in healthy, ethical, and safe digital practices.

This submission responds to the call for feedback on governance and regulatory responses, with a focus on the possible elements of a draft risk-based approach. While we will address each of the element categories, on a general level, the current proposed draft framework appears inadequate to meaningfully protect individuals and the wider community from potential AI harm. The draft framework would also be insufficient in building trust and transparency on AI development and adoption.

Overall, the current approach appears to be designed to treat AI development as part of singular industry adoption. The framework as it is currently configured does very little to address the creation of AI applications that are more generalised, and which will have wider impacts across not only whole industries and sectors, but the broader economy and society. AI governance needs to consider impacts of AI not just on automation of specific roles or tasks, but also on general rates of employment and the impact on society and the whole economy, in recognition of the profound changes which AI entails.

Enhanced regulatory oversight, with far more robust elements in a risk-based approach, would increase safety, transparency, and trust – all essential in underpinning long-term stability in Australia's development and adoption of AI. Yet beyond this, a stronger regulatory approach is also needed to ensure protections for individuals and the community. This stronger approach must address the broad current and potential impacts of AI, including on people's wellbeing and social cohesion.

For further information on any of the points raised in our submission, please contact CDW on secretariat@digitalwellbeing.org.au or 02 6162 0361.

Impact Assessments

The CDW concurs with the risk-based approach to AI development and adoption in Australia, including impact assessments considering and mitigating risk. Impact assessments can better achieve the aims of greater transparency and risk management through 1) a dedicated regulatory authority and 2) the inclusion of an unacceptable risk category which aligns with community expectations, social values, and international norms. This requires a shift in the categorisation of AI from a productivity enhancing tool, to a potential social weapon and an uncontrollable decentralised advance with no clear off switch.

A dedicated regulatory authority for AI, including for impact assessments

A regulatory authority for AI would support the implementation of objective impact assessments, providing clarity to developers and adopters, and advancing trust.

An independent authority responsible for assessments would help preclude potential market incentives to assess risk lower than actual risks. For instance, authority responsibilities would clear any confusion between assessments of creditworthiness, employability, and teacher and student performance (all medium risk in the possible draft risk management approach); with automation of discrete business processes (currently low risk). More neutral impact assessments would promote business and organisational confidence in the impact assessment process. This is especially important given the known dynamic changes in AI development, with AI types quickly proliferating and evolving. The current approach which favours self-assessments and industry monitoring its' own activities is highly likely to lead to breaches and an associated erosion of trust. A well-resourced regulatory authority would bolster trust and transparency as the industry expands.

The regulatory authority would ensure automatic peer-review in impact assessments are built into the life cycle of AI design, development and deployment. These third-party assessments would be able to make recommendations on changes during design and before the product enters the market, as well as regular reviews on safety.

The regulatory authority would align internationally with the European Union (EU) proposed governance structures and Canada's proposed new regulator (the AI and Data Commissioner). Australia's Chief Scientist Rapid Response Information Report on Generative AI noted that Canada's proposed regulator would be granted powers to require audits as well as order the suspension of an AI's system's use.[1] High-profile AI developers, such as OpenAI CEO Sam Altman, have proposed the creation of a new federal agency in the United States, and a bill recommending the establishment of a Federal Digital Platform Commission has been introduced to the U.S. Senate.[2] Such a body would be in line with submissions made to "Positioning Australia as a Leader in Digital Economy Regulation" as well as the Australian Human Rights Commission's (AHRC) 2021 *Human Rights and Technology Final Report*. The regulatory

[1] Australian Government Chief Scientist Rapid Response Information Report – Generative AI: Language models and multimodal foundation models. 24 March 2023.
[2] Brian Fung, 2023. US senator introduces bill to create a federal agency to regulate AI. CNN. 18 May 2023.

authority would build off the Australian Government's recent commitment to strengthening AI governance, supporting responsible AI deployment, and resourcing the National Artificial Intelligence Centre.[3]

Unacceptable risk category

Currently, the possible draft risk management approach does not include an explicit "unacceptable risk" category in addition to the low risk, medium risk, and high-risk categories, juxtaposed to the European Union AI Act with the unacceptable risk categorisation requiring the banning of particular AI types.

We note that one of the areas of feedback being sought are implications for Australia's domestic tech sector and trade and export activities if the government were to ban certain high-risk activities. The CDW agrees with Dr Melinda Rankin that banning AI types that pose unacceptable risks would deter the malign use of AI and ensure AI is aligned with Australia's democratic values.[4] Dr Rankin points out that rather than limiting innovation, categorisations would provide businesses and consumers with clarity on what is considered "trusted, accountable and responsible AI" – particularly important with low trust levels in AI.

Further to these points, the Australian Government's Rapid Response Information Report on Generative AI noted that the EU model has the potential to become an international standard, given it will apply to EU citizens even if systems are developed overseas. This would place any such activities in Australia deemed unacceptably risky within the EU at odds with broader partnerships and international standards.

Moreover, the CDW proposes that allowing certain high-risk or unacceptably risky AI activities to proceed would be detrimental to trust levels for all AI types. For example, AI development in Australia in areas of social scoring or practices that have significant potential to manipulate persons through subliminal techniques would likely undermine consumer trust in AI overall, including less risky AI use cases, harming Australia's domestic tech sector and AI activities.

Notices

The CDW recommends to further elaborate on notices informing users where AI is used. Notices should be easily understood by individuals, including what systems have been or are being used, how they are or may be materially affected, and how to properly seek review and redress. These notification areas will not only then help ensure transparency, but also lead to the development or refinement of appropriate complaint and feedback mechanisms responding to people adversely affected by AI systems.

Notices must also avoid the publicised issues with cookie notices seen under the EU General Data Protection Regulation (GDPR). Cookie consent notices are often used in conjunction with techniques that nudge people into accepting data collection and tracking, including notification positions, and choice offerings.[5]

[3] Australian Government, 2023. Investments to grow Australia's critical technologies industries.
[4] Melinda Rankin, 2023. Regulating artificial intelligence: How the EU just got us closer. Lowy Institute. 18 July 2023.
[5] Matt Burgess, 2020. We need to fix GDPR's biggest failure: broken cookie notices. 28 May 2020.

Human in the loop/oversight assessments

The CDW agrees with the importance of human in the loop and oversight requirements, but rather than being classified as appropriate, this risk element should be made far broader with deeper consideration of the risks of not having humans in the loop. The application should be expanded to ensure that there is at least some level of human oversight or involvement on AI applications beyond those with the lowest-risk assessments.

For example, the current possible draft risk management matrix has medium risk use cases – including preliminary assessments of business loans, student and teacher performance assessments, and hiring and employee evaluation processes – to be self-assessed by industry with human involvement commensurate with risk. This approach risks opaque systems where it is difficult for people adversely affected by AI systems to have proper understanding or access to recourse.

Indeed, the Rapid Response Information Report on Generative AI noted that: "Where an AI-enabled system has no clear human decision-makers, it is challenging – but essential – to establish responsibility for adverse impacts."[6] The report goes on to discuss how most large language models (LLMs) and multimodal foundation models (MFMs) are 'black box technologies', "where the public cannot understand how the model arrives at its outputs, making it difficult, or potentially impossible, for a human to assess the reliability of the results or seek redress."

The same would be true of many AI use cases, where the lack of appropriate levels and enforcement of human in the loop and oversight would leave people materially affected but with inappropriate means for complaints and feedback, and the inability to hold companies and developers to account.

Explanations

The CDW believes that the regulatory framework would have to extend far beyond explanations, noting their shortfalls in other areas of technological adoption. Explanations around privacy have largely failed to meaningfully protect this fundamental human right – not necessarily because people have not had access to information, but because people favour short-term convenience and usage. High profile examples include the stories about the length of time needed to read terms and conditions or privacy policies, such as the analysis that one would need 76 work days each year to read relevant internet private policies.[7] More regulatory impetus would be needed to compel developers to meaningfully clarify AI decisions for people affected by them.

[6] Australian Government Chief Scientist Rapid Response Information Report – Generative AI: Language models and multimodal foundation models. 24 March 2023.
[7] Alexis Madrigal, 2012. Reading the Privacy Policies You Encounter in a Year Would Take 76 Work Days. The Atlantic. 1 March 2012.

Training

While training is important, it is insufficient to achieve risk mitigation, oversight and supervision aims without placing further proper safeguard mechanisms in place. Breaches of regulatory and risk compliance frequently occur despite well-entrenched employee training regimes – we point to two recent examples in high-profile in the Australian banking[8] and accounting sectors. This includes cheating on online training tests "designed to improve professional safety requirements and ensure partners and staff act with integrity." [9]

Training itself must be far more than a tick-box exercise. Training completion should not act as a barometer as they "may not truly indicate learning, buy-in, or compliance to the company's policies and procedures."[10] Compliance training and programs more generally often mistake legal accountability (in training) with compliance effectiveness.[11]

Training should also not be confined to employees as proposed in the possible elements of a draft risk-based approach. Compliance efforts must be led by executives, senior leaders, and managers, ensuring that compliance is not just about training but is at the core of organisational actions, especially pertinent for trusted AI development, adoption, and regulation.

Monitoring and documentation

Closely linked to the feedback outlined in impact assessments, monitoring and documentation would benefit most from being undertaken by a dedicated regulatory authority, rather than self-monitoring and documentation. Monitoring and documentation should adhere to clear regulatory and legal requirements that are comprehensive and transparent. The framework should also include enforcement mechanisms for breaches, with proper reinforcement capability and sanctions for non-compliance, including fines and ability to suspend deleterious activities (which would align with EU and Canada's policy developments).

Much more robust regulatory capacity - with strong safeguards - would be needed for monitoring and documentation, other possible elements of the draft risk-based approach, and the overall framework — in order to ensure that AI in Australia is transparent, trusted, safe, and responsible.

[8] ASIC, Westpac penalised $113 million after multiple ASIC legal actions, 22 April 2022. Reuters, Factbox: Penalties imposed on Australian companies over recent years, 30 May 2023.
[9] Caitlin Cassidy, US watchdog fines KPMG Australia over 'widespread' cheating on online training tests. The Guardian. 15 September 2021.
[10] Deloitte, Compliance Week in Focus: 2016 Compliance Trends Surveys. 2016.
[11] Hui Chen and Eugene Soltes, Why Compliance Programs Fail – and How to Fix Them. Harvard Business Review. March – April 2018.