Australian Human Rights Commission

Mandatory Guardrails for AI in High-risk Settings

Australian Human Rights Commission

Submission to Department of Industry, Science and Resources

04 October 2024

ABN 47 996 232 602
Level 3, 175 Pitt Street, Sydney NSW 2000
GPO Box 5218, Sydney NSW 2001
General enquiries 1300 369 711
National Information Service 1300 656 419
TTY 1800 620 241

Australian Human Rights Commission
www.humanrights.gov.au

ABN 47 996 232 602
Level 3, 175 Pitt Street, Sydney NSW 2000
GPO Box 5218, Sydney NSW 2001
General enquiries 1300 369 711
National Information Service 1300 656 419
TTY 1800 620 241

1 Introduction

1. The Australian Human Rights Commission (Commission) welcomes the opportunity to make this submission to the Mandatory Guardrails for AI in High-risk Settings Proposals Paper (Proposals Paper).

2 Defining high-risk

2. A human rights risk-based approach to artificial intelligence (AI) is essential for regulation. The Commission welcomes the inclusion of human rights as a key factor in determining if AI is 'high-risk'. Given that there is currently no human rights act in Australia,[1] any classification approach will need to expressly make reference to both domestic human rights law and international human rights obligations.
3. A human rights-based approach to AI classification must be expansive to ensure that the full spectrum of rights are considered. This principles-based approach reflects the natural evolution of human rights over time.[2] It also ensures an interoperable approach with other jurisdictions incorporating human rights into their AI legislative responses.[3]

   Recommendation 1: The Federal Government should adopt a risk-based and preventative approach to AI regulation that is centred on human rights.

2.1 Principles and lists

4. The Proposals Paper presents classification of a 'principle' or 'list' based approach as alternatives to one another.[4] However, the Commission supports a combined approach. A principles-based approach is flexible and adaptable, allowing regulation to evolve and keep up with technological advancements.
5. However, businesses, developers and agencies utilising AI would also be assisted by the inclusion of a non-exhaustive list of AI applications that would also be classified as high-risk.
6. For example, the use of AI in hiring, promotion and dismissal processes remains a concern for the Commission. As previously noted, AI product may be affected by algorithmic bias that unintentionally discriminates against people based on a protected attribute.[5]

7. Compounding this, the lack of transparency and explainability in AI-generated decisions denies individuals access to redress for decisions made about their employment. The inscrutability of AI-informed decision-making risks imposing an unattainable burden of proof on people seeking remedy under the *Fair Work Act 2009* (Cth) or federal anti-discrimination legislation.[6]
8. For these reasons, the Commission believes that the inclusion of a non-exhaustive list of high-risk AI applications – such as addressing the use of AI systems to make hiring, promotional or dismissal related decisions – would be beneficial. If the use of AI in employment is not categorised as high risk, it poses a risk to an individual's right to employment on an equal basis to others, as outlined in international human rights instruments[7] or results in unlawful discrimination under domestic legislation.[8]

   Recommendation 2: The Federal Government should adopt a combined approach which utilises principles and non-exhaustive lists to classify high-risk AI.

3 Unacceptable risk

9. Some uses of AI pose an unacceptable risk to human rights and should be prohibited. For example the European Union's AI Act, specifically prohibits against AI applications presenting an 'unacceptable risk'. This includes any AI system that exploits any 'vulnerable groups' or the use of AI systems for evaluation purposes based on social behaviour or personality.[9]
10. One key example that currently poses an unacceptable risk is the use of AI in facial recognition technologies (FRT). The Commission has previously raised concerns about this use case example in both the Human Rights and Technology Project Final Report (Final Report) and past submissions.[10] A moratorium on the use of FRT in decision-making that has a legal, or similarly significant, effect for individuals (or where there is a high risk to human rights) is still needed until specific legislation to regulate this is introduced. The Commission provides in-principle support for the Human Technology Institute's Model Law on FRT to address this issue.
11. A second example of an AI application that presents an unacceptable risk is the use of deepfake sexual material. Deepfake sexual material[11] refers to deepfakes[12] that are sexual in nature. Deepfakes can be used to humiliate, extort, or silence an individual, or for sexual gratification.[13] The Federal Government recently criminalised the creation and dissemination of such

material,[14] and this approach should be reflected by listing the use of deepfake sexual material as an AI application that presents an unacceptable risk .

12. The Commission is also concerned about the potential for AI systems to be used as a tool for mass surveillance to generate social scoring systems. The collection of large amounts of information by governments or companies regarding a person's personal, financial and political conduct to create a 'social credit score' is reminiscent of the social credit system seen in China[15] and raises serious human rights concerns. Not only is such surveillance a serious breach of privacy but also inhibits a person's ability to dispute decisions made based on the collected data and restricts a range of other human rights, including freedom of expression.[16] This is another kind of AI use that should be prohibited in Australia, reflecting the prohibition in the European Union's AI Act.

Recommendation 3: The Federal Government should set out an 'unacceptable risk' category of AI uses which are prohibited.

4 Mandatory guardrails

13. Considering the unique and complex nature of harms posed by high-risk AI, it is crucial to have a robust set of guidelines that organisations can effectively adopt when engaging with AI.

14. The proposed guardrails appear to reflect the fundamental obligations being introduced or considered by numerous jurisdictions, with the most notable being the European Union's AI Act.[17] While the guardrails reflect a promising approach to AI governance, they would be further improved by strengthening the human rights protections.

4.1 Guardrails for human rights

15. The proposed guardrails do not specify the need for high-risk AI to comply with human rights obligations. This is concerning as high-risk AI can adversely affect human rights and freedoms, such as the right to privacy and non-discrimination.[18] A more direct integration of human rights into the guardrails would strengthen AI governance.

16. A human rights requirement could be included as an amendment to guardrail two in establishing and implementing a risk management process. Guardrail two already examines the potential impacts on 'people,

community groups and society'. To strengthen this safeguard, it should specify that the risk management process should consider human rights implications.

17. Where a high-risk AI product may limit people's human rights under guardrail two, developers and deployers should be required to justify the limitation on human rights (e.g. if the public interest outweighs the harm to individual human rights).

Recommendation 4: The Federal Government amend guardrail two to expressly include human rights considerations.

4.2 Transparency and explicability

18. Guardrail seven is crucial to ensure that people can contest AI-informed decisions or make complaints about their experience or treatment.[19] While organisations are required to provide 'sufficient information about the use or outcomes of the AI system' that information will be redundant unless it is also understandable and accessible.

19. A prevalent issue with challenging AI-informed outcomes is the difficulty understanding the functions of system and its algorithm. Known as the 'black box' phenomenon, this issue makes it difficult for contesters to prove harmful impacts of systems affected by bias.[20]

Recommendation 5: The Federal Government amend guardrail seven to require provided information to be understandable and accessible.

5 Regulatory options

20. Each of the proposed options for introducing mandatory guardrails come with benefits and limitations. Considering the need for a consistent approach to AI governance, the Commission supports option three of the Proposals Paper which would create an Australian AI Act.[21]

Recommendation 6: The Federal Government should pursue option three of the Proposals Paper and introduce an Artificial Intelligence Act.

21. However, the introduction of an Australian AI Act is neither a panacea nor a timely solution to the most pressing issues posed by AI. There remains a need to urgently address specific examples of harm that have arisen due to new and emerging AI tools. One example of necessary reform has been the passage of the *Criminal Code Amendment (Deepfake Sexual Material) Bill*

2024 (Cth) which criminalises the creation and transmission of deepfake pornography. It is these kinds of reforms which should not be delayed pending the development of an AI Act.

22. The introduction of an AI Act is only part of the solution. The Federal Government must continue to address the most urgent harmful impacts of AI through ongoing law reform efforts (in addition to creating an AI Act).

5.1 Urgency of reform

23. Irrespective of which regulatory option the Federal Government pursues, mandatory guardrails are urgently required to mitigate the human rights risks associated with AI.

24. The Commission directly outlined the risks of AI and the need for a legislative response to the Federal Government approximately three years ago. In 2021 the Commission published its Final Report outlining 38 recommendations to ensure that AI is developed and deployed ethically in Australia. The majority of these recommendations have not been implemented, despite the Final Report offering a proactive response to ensure the early development of AI in Australia in responsible and ethical ways.

25. While the consultative nature of both the Discussion Paper and Proposals Paper are welcomed, the pace of reform in Australia is failing to keep up with the rapid development of AI. Clear and accurate timelines for the introduction of AI regulation must be provided to ensure accountability and timeliness.

Recommendation 7: The Federal Government must provide clear and accurate timelines for the introduction of AI regulation.

5.2 AI Commissioner

26. The Proposals Paper seeks to establish an independent AI regulator to oversee a 'monitoring and enforcement regime'.[22] While a positive initiative, limiting a regulator to performing monitoring and enforcement functions is insufficient for the reasons set out below.

27. Businesses, government agencies and the broader public are adopting AI rapidly – often without understanding the risks or upcoming regulatory implications. Education will be a key factor to both ensure compliance with AI regulation and the ethical development and deployment of AI.

28. The Commission has for some time called for the creation of an AI Commissioner as an independent statutory office.[23] In addition to providing expert advice on how to comply with laws and ethical standards that apply to the development and use of AI, the Commissioner could play a key role in building the capacity of existing regulators to adapt and respond to the rise of AI.[24]
29. As an independent statutory office that champions public interests, an AI Commissioner could have a critical role in advancing public interest, awareness and education in the safe use of AI.[25] As a result, an AI Commissioner can have a meaningful and expansive role in the governance of AI law reform, enforcement and community engagement.

    Recommendation 8: The Federal Government establish an independent statutory AI Commissioner as proposed by the Australian Human Rights Commission in its 2021 Final Report.

6 Recommendations

30. The Commission makes the following recommendations.

    Recommendation 1:

    The Federal Government should adopt a risk-based and preventative approach to AI regulation that is centred on human rights.

    Recommendation 2:

    The Federal Government should adopt a combined approach which utilises principles and non-exhaustive lists to classify high-risk AI.

    Recommendation 3:

    The Federal Government should set out an 'unacceptable risk' category of AI uses which are prohibited.

    Recommendation 4:

    The Federal Government should amend guardrail two to expressly include human rights considerations.

    Recommendation 5:

    The Federal Government should amend guardrail seven to require provided information to be understandable and accessible.

Recommendation 6:

The Federal Government should pursue option three of the Proposals Paper and introduce an Artificial Intelligence Act.

Recommendation 7:

The Federal Government must provide clear and accurate timelines for the introduction of AI regulation.

Recommendation 8:

The Federal Government should establish an independent statutory AI Commissioner as proposed by the Australian Human Rights Commission in its 2021 Final Report.

Endnotes

[1] The Australian Human Rights Commission has proposed a national human rights act, which was recently supported by Parliamentary Joint Committee on Human Rights, *Inquiry into Australia's Human Rights Framework* (Report, May 2024).

[2] For example, a healthy, clean and sustainable environment has in recent years been recognised as a universal human right. UNHRC, *Resolution 48/13: The human right to a clean, healthy and sustainable environment* (UN Doc. A/HRC/RES/48/13, 8 October 2021); UNGA, *The Human Right to a Clean, Healthy and Sustainable Environment* (resolution A/76/L.75, 01 August 2022).

[3] See e.g. *Regulation (EU) 2024.1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence* [2024] OJ L 1689; Government of Canada, *The Artificial Intelligence and Data Act (AIDA) – Companion document* (Webpage) <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>.

[4] Department of Industry, Science and Resources, *Mandatory Guardrails for AI in High-risk Settings* (Proposals Paper, 05 September 2024) 26-27.

[5] AHRC, Submission No 212 to the Department of Industry, Science and Resources' *Supporting Responsible AI: Discussion Paper* (Submission, 26 July 2023) 31-32.

[6] *Sex Discrimination Act (No 4) 1984* (Cth); *Racial Discrimination Act (No 52) 1975* (Cth); *Disability Discrimination Act (No 135) 1992* (Cth); *Age Discrimination Act (No 68) 2004* (Cth).

[7] *Convention on the Rights of Persons with Disabilities*, art 27; *Convention on the Elimination of all Forms of Discrimination Against Women*, art 11; *International Convention on the Elimination of All Forms of Racial Discrimination* art 5(i); *International Covenant on Economic, Social and Cultural Rights*, art 7(b).

[8] See e.g. *Racial Discrimination Act 1975* (Cth); *Sex Discrimination Act 1984* (Cth); *Disability Discrimination Act 1995* (Cth); *Age Discrimination Act 2004* (Cth).

[9] *Regulation (EU) 2024.1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence* [2024] OJ L 1689 art 5.
[10] AHRC, *Final Report* (Final Report, 2021) recs 19 & 20; see also AHRC, Submission No 212 to the Department of Industry, Science and Resources' *Supporting Responsible AI: Discussion Paper* (Submission, 26 July 2023) 43-49.
[11] Sometimes known as 'deepfake pornography', or a form of 'intimate image or video'.
[12] A deepfake is a 'digital photo, video or sound file of a real person that has been edited to create an extremely realistic but false depiction of them doing or saying something that they did not actually do or say'; eSafety Commissioner, *Deepfake Trends and Challenges – Position Statement* (Webpage) <https://www.esafety.gov.au/industry/tech-trends-and-challenges/deepfakes>.
[13] Stephanie Tong, '"You Won't Believe What She Does": An Examination into the Use of Pornographic Deepfakes as a Method of Sexual Abuse and the Legal Protections Available to its Victims' (2022) 22-25 *UNSW Law Journal Student Series*, <https://www5.austlii.edu.au/au/journals/UNSWLawJlStuS/2022/25.html>; Danielle K. Citron and Robert Chesney, 'Deep Fakes: A Looming Challenge for Privacy, Democracy and National Security' (2019) 107 *California Law Review* 1753, 1772.
[14] See generally *Criminal Code Amendment (Deepfake Sexual Material) Bill 2024* (Cth).
[15] Xu Xu, Genia Kostka and Xun Cao, 'Information Control and Public Support for Social Credit Systems in China' (2022) 84(4) *Journal of Politics* 2230, 2230.
[16] Simon Burgess and Matthew Wysel, 'China's Social Credit System: How Robust is the Human Rights Critique' in *Who's Watching? Surveillance, Big Data and Applied Ethics in the Digital Age* (2022) vol 26, 39-40.
[17] *Regulation (EU) 2024.1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence* [2024] OJ L 1689.
[18] See generally, Volker Turk, 'Artificial intelligence must be grounded in human rights, says High Commissioner' (Statement, 12 July 2023); The significant human rights harms were also AHRC, Submission No 212 to the Department of Industry, Science and Resources' *Supporting Responsible AI: Discussion Paper* (Submission, 26 July 2023).
[19] AHRC, Submission No 212 to the Department of Industry, Science and Resources' *Supporting Responsible AI: Discussion Paper* (Submission, 26 July 2023) 37.
[20] AHRC, Final Report (Final Report, 2021) 65.
[21] AHRC, Submission No 212 to the Department of Industry, Science and Resources' *Supporting Responsible AI: Discussion Paper* (Submission, 26 July 2023) 33.
[22] Department of Industry, Science and Resources, *Mandatory Guardrails for AI in High-risk Settings* (Proposals Paper, 05 September 2024) 49.
[23] AHRC, *Final Report* (Final Report, 2021) 125; AHRC, Submission No 212 to the Department of Industry, Science and Resources' *Supporting Responsible AI: Discussion Paper* (Submission, 26 July 2023) 35-36.
[24] AHRC, *Final Report* (Final Report, 2021) 125; AHRC, Submission No 212 to the Department of Industry, Science and Resources' *Supporting Responsible AI: Discussion Paper* (Submission, 26 July 2023) 34.
[25] AHRC, *Final Report* (Final Report, 2021) 125; AHRC, Submission No 212 to the Department of Industry, Science and Resources' *Supporting Responsible AI: Discussion Paper* (Submission, 26 July 2023) 35.