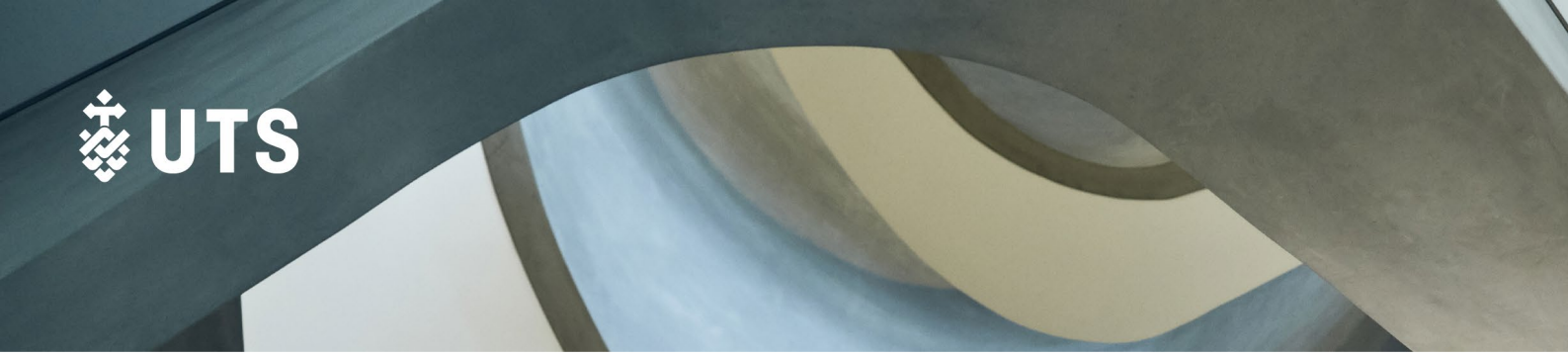


Consultation on Safe and Responsible AI in Australia

Dr Ramona Vijeyarasa, Associate Professor, Faculty of Law,

University of Technology Sydney

Introduction

This submission is focused on providing a gender perspective to how the Australian Government should develop its regulatory and governance response to the design, deployment and use of AI-driven technologies. I write this submission as one of Australia's key academic experts on gender and the law. As an Associate Professor in the Faculty of Law at the University of Technology Sydney, I bring to my academic experience 10 years of practical experience as a women's human rights lawyer. I have worked in local and international non-government and inter-governmental organisations in jurisdictions across the globe. I have published three books and dozens of articles on bringing a gender perspective to law-making. I am the chief investigator behind the research collaboration that led to the Gender Legislative Index, an online tool that uses human evaluators and machine learning to assess the gender-responsiveness of legislation. In 2022, I was named the Woman in Artificial Intelligence (Law category) in Australia & New Zealand and 2nd Runner-Up for the Woman in AI Innovator of the Year award.

Recommendations

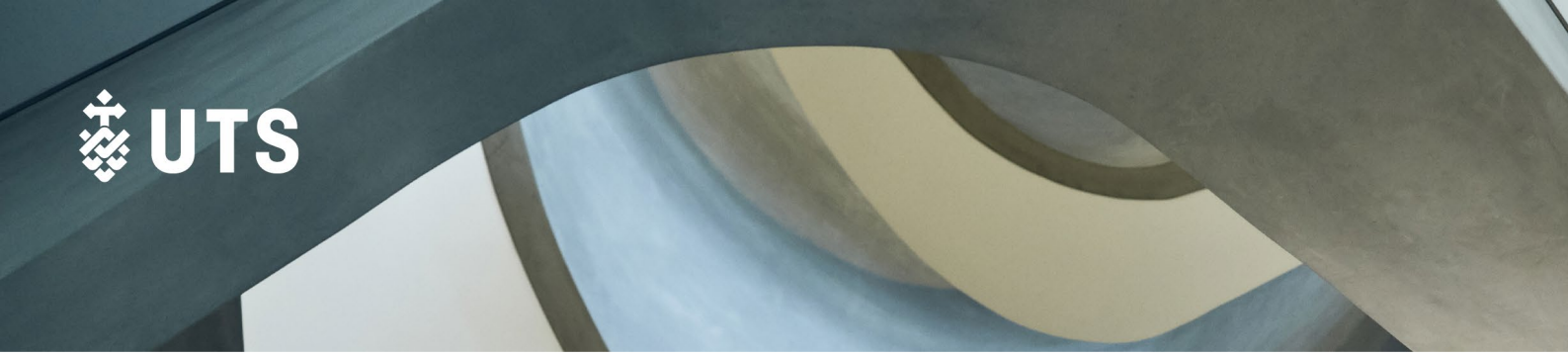
This submission particularly focuses on Questions 2, 3 and 5 of the call for contributions. I propose three concrete recommendations for Australia to bring a gender-perspective to its regulatory response to AI.

Recommended approaches to regulation

1. **Adopt a gender lens to any future Australian Artificial Intelligence Act**, by defining and incorporating into the law principles of non-discrimination including on the basis of gender and sexual orientation and accountability for risks concerning disadvantaged and vulnerable groups (Terms of Reference, Q3, Q4, Q5)
2. **Strengthen the accountability of entities designing, deploying or using AI** through rigorous requirements for gender assessments and independent mechanisms to flag harmful content (Terms of Reference Q5)
3. **Ensure the active participation of a diversity of women in the governance of AI** by incorporating and defining gender-sensitive due diligence and a requirement for responsible data collection (Terms of Reference Q3)

Detailed recommendations:

1. **Adopt a gender lens to any future Australian Artificial Intelligence Act, by defining and incorporating into the law principles of non-discrimination including on the basis of gender and sexual orientation and risks concerning disadvantaged and vulnerable groups (Terms of Reference Questions 1c,1e)**



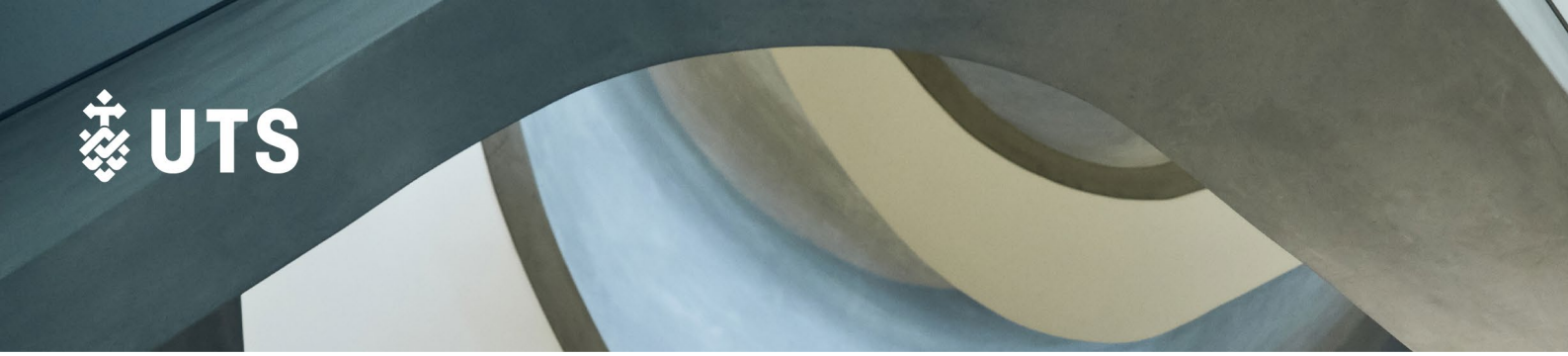
The gendered harms of AI are increasing well-known and have been the subject of a significant body of research in this field.[i] As a gendered problem, our legislative response must be gender-responsive, that is, the law must take into account differences in interests, needs and experiences of men, women and non-binary people.[ii] A gender-responsive approach to addressing the harms of AI involves recognizing what can broadly be categorised as three types of harms:

a) **Allocative harms** result from decisions about how to allocate goods and opportunities among a group. Here we can think of the way an AI system used in a recruitment process may disproportionately classify applications for male candidates as more suitable than female. Such a system potentially results in a loss of financial opportunities, livelihoods and freedom of choice for women when compared to men.[iii]

b) **Representational harm** comes about when systems reinforce gendered subordination, through stereotyping, under-representation or denigration. Numerous examples exist that reify heteronormative gender roles and objectify women in the process. A commonly-cited example is the use of female voices in AI-powered virtual assistants – Amazon's Alexa, Apple's Siri, Microsoft's Cortana and Google's Voice Assistant. A further example is the use of 'deepfakes' in pornography-related attacks that are a *non-consensual* form of gender-based online abuse.

c) **Knowledge-based harms account** for the inequalities that exist concerning the different levels of understanding regarding how algorithms influence everyday lives and the different skills and creative techniques needed to re-correct algorithms' biases.[iv] This inequality comes about and is perpetuated by the over-representation of men in AI's design, deployment, use and overarching dominance of leadership in the technology space, with inadequate and unequal representation of women.

Significantly less attention has been drawn to the specific impacts of AI on lesbian, gay, bisexual, transgender, and queer (LGBTQ+) communities, particularly given that surveillance technologies are incapable of working beyond the binaries of male and female.[v] Indeed, there is a far greater complexity to the discussion on gender and AI than I am able to do justice to in this submission, if we think about the need to go beyond binaries in the use of automated body scanners, facial recognition or social media content filtering, just to name a few examples. In this context, the regulatory proposals I suggest in this submission have limitations; when it comes to gender diversity, more emphasis may be needed on design-oriented solutions, with a strong focus on self-determination by AI's users and better acknowledging gender pluralism in AI's design.[vi]



In order to regulate in response to the known and potential gendered harms of AI-driven technology, Australia can benefit from the regulatory approaches currently debated in **Canada** in its treatment of biased output (prohibiting discrimination while permitting positive discrimination through the use of AI-driven technologies). Australia can also benefit from **Brazil's** latest draft bill and its approach to non-discrimination and the disproportionate impact of AI on vulnerable and specific groups.

Canada's *Artificial Intelligence and Data Bill, Bill C-27 (An Act to Enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to Make Consequential and Related Amendments to Other Acts*)

biased output means content that is generated, or a decision, recommendation or prediction that is made, by an artificial intelligence system and that adversely differentiates, directly or indirectly and without justification, in relation to an individual on one or more of the prohibited grounds of discrimination set out in section 3 of the *Canadian Human Rights Act*, or on a combination of such prohibited grounds. It does not include content, or a decision, recommendation or prediction, the purpose and effect of which are to prevent disadvantages that are likely to be suffered by, or to eliminate or reduce disadvantages that are suffered by, any group of individuals when those disadvantages would be based on or related to the prohibited grounds. (*résultat biaisé*)[1]

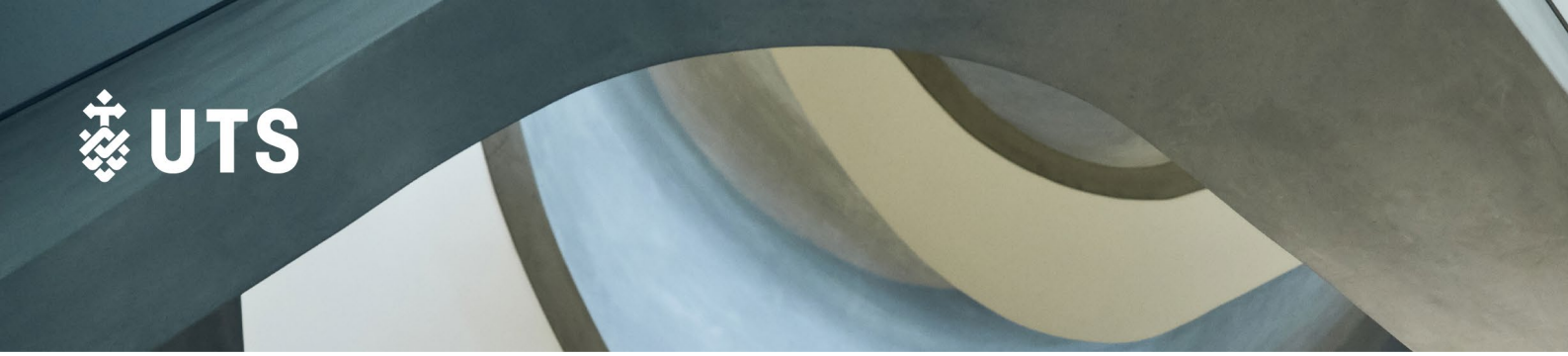
Brazil, Bill N° 2338/2023 Bill for the Use of Artificial Intelligence (Dispõe sobre o uso da Inteligência Artificial) [unofficial translation]

Art. 12.

Persons affected by decisions, forecasts or recommendations made by artificial intelligence systems are entitled to fair treatment and a prohibition on the implementation and use of artificial intelligence systems that may lead to direct, indirect, illegal or abusive discrimination, including:

(i) as a result of the use of sensitive personal data or disproportionate impacts due to personal characteristics such as geographic origin, race, colour or ethnicity, gender, sexual orientation, socioeconomic class, age, disability, religion or political opinions;

(ii) due to the establishment of disadvantages or aggravation of the situation of vulnerability of people belonging to a specific group, even if apparently neutral criteria are used.



2. Strengthen the accountability of entities designing, deploying or using AI through rigorous requirements for gender assessments (Terms of Reference Q5)

a) Undertaking gender-specific impact assessments

As early as 2019, Canadian introduced the T*reasury Board Directive on Automated Decision-Making*, a mandatory policy instrument which applies to almost all federal government institutions.[vii] It places a particular emphasis on impact assessments and transparency, including the likely impacts of AI on freedom, health, the economy and environment.[viii]

The standout feature of the Canadian Directive is the Gender-Based Analysis Plus, a quality reassurance requirement before launching into production of an AI-driven technology.[ix] This Gender-Based Analysis Plus requirement entails compulsory testing for unintended biases. If an AI-driven technology meets the moderate, high and very high-risk thresholds, the designers need to undertake Gender-Based Analysis Plus. This "plus" reflects going further than the gender impact assessment already required for procurement. Additional steps include an assessment of the impact of the automation on gender and/or other identifying factors but also naming what planned or existing measures are in place to address these identified risks in the future.[x]

Canadian Gender-Based Analysis Plus 2022 (specific to AI)
Ensure that the Gender-based Analysis Plus addresses the following issues:

- impacts of the automation project (including the system, data and decision) on gender and/or other identity factors;
- planned or existing measures to address risks identified through the Gender-based Analysis Plus.

b) Flagging harmful online content

Australia can learn from the EU's approach to addressing harmful online content. The *EU Digital Services Act* has been in force since 2022. In response to the issue of deepfakes and harmful AI-driven online content, the *EU Digital Services Act* offers the figure of the **trusted flagger**, effectively a form of appointed, expert and independent industry whistleblower.[xi] Particularly important is the fact that trust flaggers are independent of the provider/platform.

Given the particular harms facing women and girls from deepfakes, the figure of the whistleblower offers a gender-sensitive response to an evident gender-based harm from AI. Organisations are appointed in the role of 'trusted flagger' if they meet predefined criteria such as a specific expertise in illegal content, independence from the platforms and integrity of its activities. Government funding is received for the role. Trusted Flaggers should publish reports at least once a year on their activities. Once content is flagged, the responsible entity – such as META, or in the case of cloud servers such as drop box – would be required to remove the content; if the content is not removed, the entity must explain why.[xii]

EU Digital Services Act 2022 (Article 22) Trusted flaggers

1. Providers of online platforms shall take the necessary technical and organisational measures to ensure that notices submitted by trusted flaggers, acting within their designated area of expertise, through the mechanisms referred to in Article 16, are given priority and are processed and decided upon without undue delay.
2. The status of 'trusted flagger' under this Regulation shall be awarded, upon application by any entity, by the Digital Services Coordinator of the Member State in which the applicant is established, to an applicant that has demonstrated that it meets all of the following conditions:
   a) it has particular expertise and competence for the purposes of detecting, identifying and notifying illegal content;
   b) it is independent from any provider of online platforms;
   c) it carries out its activities for the purposes of submitting notices diligently, accurately and objectively.
3. Trusted flaggers shall publish, at least once a year easily comprehensible and detailed reports on notices submitted in accordance with Article 16 during the relevant period. The report shall list at least the number of notices categorised by:
   a) the identity of the provider of hosting services,
   b) the type of allegedly illegal content notified,
   c) the action taken by the provider.

Those reports shall include an explanation of the procedures in place to ensure that the trusted flagger retains its independence. Trusted flaggers shall send those reports to the awarding Digital Services Coordinator, and shall make them publicly available. The information in those reports shall not contain personal data.

3. Trusted flaggers shall publish, at least once a year easily comprehensible and detailed reports on notices submitted in accordance with Article 16 during the relevant period. The report shall list at least the number of notices categorised by:

(a) the identity of the provider of hosting services,
(b) the type of allegedly illegal content notified,
(c) the action taken by the provider.

Those reports shall include an explanation of the procedures in place to ensure that the trusted flagger retains its independence. Trusted flaggers shall send those reports to the awarding Digital Services Coordinator, and shall make them publicly available. The information in those reports shall not contain personal data.

3. Ensure the active participation of a diversity of women in the governance of AI by incorporating and defining gender-sensitive due diligence and a requirement for responsible data collection (Terms of Reference Q3)

The over-representation of men in the design of AI-driven technologies is widely acknowledged. Meanwhile women dominate among those scholars identifying the gender-based biases that are and can result from AI's deployment. In turn, the presence of a greater diversity of views to represent the interests of those affected by such biases is arguably an important part of the solution.[xiii] This is an argument increasingly made,[xiv] that is, a higher participation of women in AI-driven technologies is needed to bring visibility to the gender-based harms of AI.

Non-regulatory requirements for AI should include an obligation on private and public entities deploying AI-driven technologies to conduct gender-sensitive due diligence, ensuring the active participation of a diversity of women in AI's governance.

Gender-sensitive due diligence places a particular emphasis on the experiences of women and girls, and the multiple intersecting forms of discrimination that influence the realisation of equal rights.[xv] Such gender-responsive due diligence requires, as a starting point, recognition of the embedded gender norms, complex **gender, socio-economic and** cultural **and racial** biases **at play** and power imbalances involved in **the design, deployment and use of AI-driven technologies.**

This may take the form of ensuring that individual women and women's rights organisations are given the task of the trusted whistleblower. It might be encouraging gender quotas and targets among the most dominant developers of AI-technologies as part of their corporate social responsibility. The promotion of women in such organisations to leadership roles where they have a voice and say in decision-making should also be encouraged.

It is recommended that the following definitions be incorporated into the regulatory response:

gender-sensitive due diligence means: Meaningful engagement with women and girls as relevant stakeholders, in order to understand their concrete experiences of AI-driven technologies and any adverse impacts on them.

responsible data collection means: Collection of data in order to give attention to how different groups are affected by AI-driven technologies, including but not limited to, on the basis of sex, indigenous status, racial, ethnic and sexual minority status, women and girls living with disabilities, adolescents, older women, unmarried women, women heads of household, widows, women and girls living in poverty in both rural and urban settings, women in sex work and migrant women.

Dr Ramona Vijeyarasa

Associate Professor, Faculty of Law, University of Technology Sydney

26 July 2023

[i] José-Miguel Bello y Villarino and Ramona Vijeyarasa, 'International Human Rights, Artificial Intelligence and the Challenge for the Pondering State: Time to Regulate?' [2022] *Nordic Journal of Human Rights*; Elizabeth Coombs and Halefom Abraha, *Governance of AI and Gender: Building on International Human Rights Law and Relevant Regional Frameworks* (2022) ('*Governance of AI and Gender*'); United Nations Educational, Scientific and Cultural Organization, *Artificial Intelligence and Gender Equality: Key Findings of UNESCO's Global Dialogue - UNESCO Digital Library* (UNESCO, 2020) <https://unesdoc.unesco.org/ark:/48223/pf0000374174>; Susan Leavy, 'Gender Bias in Artificial Intelligence: The Need for Diversity and Gender Theory in Machine Learning' in *2018 IEEE/ACM 1st International Workshop on Gender Equality in Software Engineering (GE)* (2018) 14 ('Gender Bias in Artificial Intelligence'); Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification' in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (PMLR, 2018) 77 <https://proceedings.mlr.press/v81/buolamwini18a.html> ('Gender Shades').

[ii] Ramona Vijeyarasa, 'Women, Work and Global Supply Chains: The Gender-Blind Nature of Australia's Modern Slavery Regulatory Regime' (2020) 26(1) *Australian Journal of Human Rights* 74, 6.

[iii] 'What Are the Accountability and Governance Implications of AI?' (21 March 2023) <https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/guidance-on-ai-and-data-protection/what-are-the-accountability-and-governance-implications-of-ai/>.

[iv] Massimo Ragnedda, 'New Digital Inequalities. Algorithms Divide' in Massimo Ragnedda (ed), *Enhancing Digital Equity: Connecting the Digital Underclass* (Springer International Publishing, 2020) 61, 61 <https://doi.org/10.1007/978-3-030-49079-9_4>.

[v] Katyal and Jung, *supra* note 56.

[vi] Katyal and Jung, *supra* note 56 at 762–763.

[vii] Government of Canada, *Treasury Board Directive on Automated Decision-Making*, (2019), https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592.

[viii] Government of Canada, *Algorithmic Impact Assessment Tool*, (2023), https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html.

[ix] Government of Canada, *Directive on Automated Decision-Making*, 6.3.6 (2019), https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592 (last visited Jun 16, 2023).

[x] Women and Gender Equality Canada, *Gender-Based Analysis Plus (GBA Plus)*, (2021), https://women-gender-equality.canada.ca/en/gender-based-analysis-plus.html (last visited Jun 16, 2023); Government of Canada, *supra* note 179 at Appendix C, 6.3.6.

[xi] European Parliament and Council of the European Union, *Digital Services Act*, REGULATION (EU) 2022/2065, DSA 61 and 62 (2022), https://eur-lex.europa.eu/eli/reg/2022/2065/oj.

[xii] *Id.* at 6(1)(b).

[xiii] Leavy (n 31) 14.

[xiv] Jackson (n 8) 316.

[xv] Office of the High Commissioner for Human Rights, *Working Paper - Gender-Sensitive Human Rights Due Diligence* (7th UN Forum on Business and Human Rights, 2018).