POLICY BRIEF

SAFEGUARDING WOMEN'S RIGHTS IN THE AI ERA: A CALL FOR GENDER-RESPONSIVE POLICIES

Artificial Intelligence (AI) is rapidly reshaping our societies - from hiring decisions and healthcare diagnostics to content recommendations and digital security. Yet, as it becomes more embedded in everyday life, AI systems are increasingly reproducing and amplifying harmful gender biases. This policy brief advocates for regulatory frameworks to ensure AI development and usage is fair, transparent, and gender-inclusive.

What's at stake?

Bias in AI is not just a technical flaw - it is a human rights issue. It stems from skewed datasets, flawed model design, and the underrepresentation of women in AI development. The result? Automated systems that consistently deliver unequal outcomes for women and marginalised groups.

A study at Berkeley Haas Center identified 133 AI systems exhibiting bias, with over 44% showing gender bias and 26% combining gender and racial bias. These systems aren't just unfair - they are unsafe. These systems frequently erase marginalised identities, especially in facial recognition technologies, pose health risks, including misdiagnoses in skin cancer detection for Black women and deliver lower-quality services to women and non-binary individuals as in voice recognition failures.[1]

But the problem runs deeper than data. AI is shaping how we think. A cognitive study revealed that even brief exposure to biased search results was enough to alter participants' perceptions about gender roles - underscoring how algorithmic bias reinforces social inequality.[2]

Without urgent action, AI will not only mirror our existing inequalities - it will magnify them. The risk is not just technological failure. It is the erosion of hard-won gains in gender equality, dignity, and human rights.

SUMMARY POLICY RECOMMENDATIONS

1.Promote Global AI Frameworks: Encourage the adoption of international AI frameworks rooted in human rights to ensure a consistent and safe global approach.

2.Support Binding Agreements on AI and Equality: Advance international and regional legal instruments that address gender-based and intersectional harms in AI systems.

3.Ensure Inclusive Global Participation: Establish fair and diverse global forums where all states and stakeholders can shape AI governance.

4.Strengthen Accountability in AI Systems: Require companies and institutions to assess and address bias in AI systems, especially in areas impacting human rights.

5.Increase Diversity in AI Development: Promote gender equality and inclusion in AI teams to foster more just technologies.

Research Approach

This policy paper is based on findings from the diploma thesis 'The Impact of Artificial Intelligence on Women's Human Rights: An International Legal Analysis'[3]. The research applies an interdisciplinary legal approach, combining international human rights law with insights from AI ethics and technological development. It draws on an extensive review of legal texts, policy documents, academic literature, and real-world examples of gender-biased AI systems. While the analysis focuses primarily on the binary gender perspective due to data limitations, it underscores the broader risks AI poses to gender equality and highlights urgent regulatory gaps.

Examples

1. Amazon's Biased Recruiting Tool

Amazon discontinued an AI recruiting tool after discovering it downgraded CVs containing the word "women's", such as "women's chess club captain". The tool was trained on past hiring patterns in the male-dominated tech sector, leading it to prefer male-coded language like "executed" or "captured", and penalise applicants with indicators linked to women.[4]

⇨ Key point: AI trained on biased historical data can systematically disadvantage women in hiring.

2. Gender Bias in Social Media Algorithms

Image moderation systems on platforms like Instagram and LinkedIn were found to classify women's bodies as more sexually suggestive than men's. In one test, Microsoft's algorithm rated a woman's photo as 96% suggestive, while a similar photo of two men received only 14%. The women's post was likely 'shadowbanned', receiving just 8 views versus 655 for the men's.[5]

⇨ Key point: AI-based content moderation can lead to disproportionate censorship and objectification of women.

3. Sexism in Word Embedding Models

Word embeddings, such as those used in search engines or chatbots, map meanings between words. A study found that these models linked "man" to "computer programmer" and "woman" to "homemaker". Similarly, analogies like "man is to doctor as woman is to nurse" emerged, revealing deep-seated gender bias even in professionally written training texts.[6]

⇨ Key point: Language-based AI tools can encode and reinforce gender stereotypes even at a structural level.

Key Findings

1.AI Systems Can Reinforce and Reproduce Discrimination

AI technologies are often trained on biased historical data and shaped by unequal social structures, resulting in both direct and indirect discrimination against women. Examples include:

- Recruitment algorithms that penalise resumes referencing women's experiences.
- Ad delivery systems on platforms like Facebook that disproportionately exclude women from high-paying job opportunities.
- Credit scoring tools that assign lower creditworthiness to women despite identical or better financial profiles.

2. Gender Stereotypes Are Embedded in AI Outputs

Language models, image tools, translation systems, and digital assistants perpetuate harmful gender norms:

- Virtual assistants with passive "female" personas respond inadequately to harassment.
- Image-generating tools oversexualise women and reinforce traditional gender roles.
- Translation and word embedding systems default to male pronouns in professional contexts, linking men with leadership and women with caregiving roles.

3. Intersectional and Structural Biases Are Amplified

Facial recognition systems and medical AI tools exhibit higher error rates for women, particularly Black women, leading to serious real-world consequences in areas such as healthcare, public safety, and surveillance. These are clear cases of intersectional discrimination, combining race, gender, and other social factors.[7]

Results

- **Existing Legal Frameworks Lack AI-Specific Provisions:** International human rights instruments like CEDAW provide crucial protections against discrimination but fail to address the unique challenges posed by AI, especially regarding gender bias in AI systems.
- **AI-Specific Regulations Are Still In Development:** While the EU AI Act and the Council of Europe's Convention on AI provide some regulation, they are geographically limited and not universally applicable. The Convention on AI, for instance, is binding only for signatory states, leaving many countries without enforceable protections.
- **Global Participation Remains Inconsistent:** The UN's resolution on AI emphasises cooperation and calls for gender equality, but lacks binding obligations, making it more symbolic than effective. As a result, global participation remains fragmented, with many states excluded from the discussions and protections.
- **Need for Binding Global Regulations:** To effectively address AI's risks, including gender bias, global, enforceable legal frameworks are needed. Current efforts lack binding mechanisms and do not apply universally. A coordinated international approach is necessary to ensure comprehensive protection for women's human rights in the AI landscape.[8]

Policy Recommendations

1.Promote the Development of Global Frameworks on AI Governance
Without worldwide commitment, AI governance remains fragmented, creating regulatory blind spots that allow harmful practices to persist. Ensuring every country engages in some form of AI regulation is vital for global safety, fairness, and consistency.

⇨ Encourage the widespread adoption of international frameworks aligned with human rights principles, emphasising gender equality and non-discrimination, to foster a globally coherent approach to AI safety.

2. Advance Binding Agreements on AI and Gender Equality Where Feasible
Many existing initiatives, such as the UN Resolution on AI, lack binding obligations and therefore have limited enforcement power. Still, global consensus on binding rules remains difficult.

⇨ Promote the negotiation of binding international agreements on AI and human rights, while also supporting regional and national-level legal instruments where global consensus is not yet possible. These should specifically address gender-based and intersectional harms.

3. Foster Inclusive Global Participation in AI Governance Debates
Current AI policy discussions are dominated by a handful of powerful countries and regions, sidelining the voices of other states and civil society actors. This lack of inclusivity undermines both fairness and effectiveness.

⇨ Establish global forums that ensure equal participation of all states, particularly those currently underrepresented, and promote gender-balanced, cross-sectoral representation in AI standard-setting bodies.

4. Improve Transparency and Accountability in AI Systems
AI systems often reflect and reinforce unfair biases, especially when they're trained on data that already contains discrimination. Without proper checks, this can lead to serious harm - especially for marginalised groups.

⇨ Companies and public institutions should be required to carefully review their AI systems to make sure they don't reproduce discrimination. This includes testing for bias and taking action when problems are found, particularly in areas where AI decisions can affect people's rights and opportunities.

5. Promote Gender Equality and Diversity in Teams
AI systems reflect the values and perspectives of the people who create them. When development teams lack diversity, there's a higher risk that systems will overlook or misunderstand the needs of marginalised groups.

⇨ Support policies that encourage the inclusion of women and underrepresented groups in AI-related education, research, and industry. Diverse teams lead to more inclusive technologies and help prevent the reproduction of harmful biases.

References
[1] Genevieve Smith and Ishita Rustagi, 'When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity' Stanford Social Innovation Review (31 March 2021) <ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity#>.
[2] Madalina Vlasceanu and David M. Amodio, 'Propagation of societal gender inequality by internet search algorighms' (12 July 2022) 119, 29 Proceedings of the National Academy of Sciences (PNAS), 5 <doi.org/10.1073/pnas.2204529119>.
[3] Stephanie Grasser, 'The Impact of Artificial Intelligence on Women's Human Rights: An International Legal Analysis' (2024), <netidee.at/impact-artificial-intelligence-womens-human-rights>.
[4] Jeffrey Dastin, 'Amazon scraps secret AI recruiting tool that showed bias against women' Reuters (11 October 2018) <reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>.
[5] Gianluca Mauro and Hilke Schellmann, ''There is no standard': investigation finds AI algorithms objectify women's bodies' The Guardian (8 February 2023) <theguardian.com/technology/2023/feb/08/biased-ai-algorithms-racy-women-bodies>.
[6] Tolga Bolukbasi and others, 'Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings' (2016) NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems, 1f <doi.org/10.48550/arXiv.1607.06520>.
[7] Further information in Chapter IV of the thesis.
[8] Further information in Chapter V of the thesis.

Policy brief written by Stephanie Grasser, based on the diploma thesis "*Impact of Artificial Intelligence on Women's Human Rights: An International Legal Analysis"*, submitted in completion of the Master's in Law at University of Graz under the academic supervision of Ao.Univ.-Prof. Mag. Dr.iur. Gerd Oberleitner.