unesco

“I don’t have a gender, consciousness, or emotions. I’m just a machine learning model.”

unesco
Centre
Under the auspices
of UNESCO

IRC AI
International Research Centre
on Artificial Intelligence
under the auspices of UNESCO

An introduction to a forthcoming Gender bias in Artificial Intelligence report coming out on March 8, 2024.

CALL TO ACTION

Share your opinion and fill in the form Public Consultation on Essay on Gender and AI

CLICK TO SUBMIT

As we stand on the precipice of a technological revolution driven by Artificial Intelligence (AI), it is imperative to ensure that this future is shaped equitably, representing all genders. With this essay we are excited to announce our forthcoming in-depth report on Gender and Artificial Intelligence in a partnership between IRCAI and UNESCO, set for release on March 8, 2024. As we prepare for this milestone event, we extend an invitation to experts, scholars, and all interested stakeholders to join us in our research.

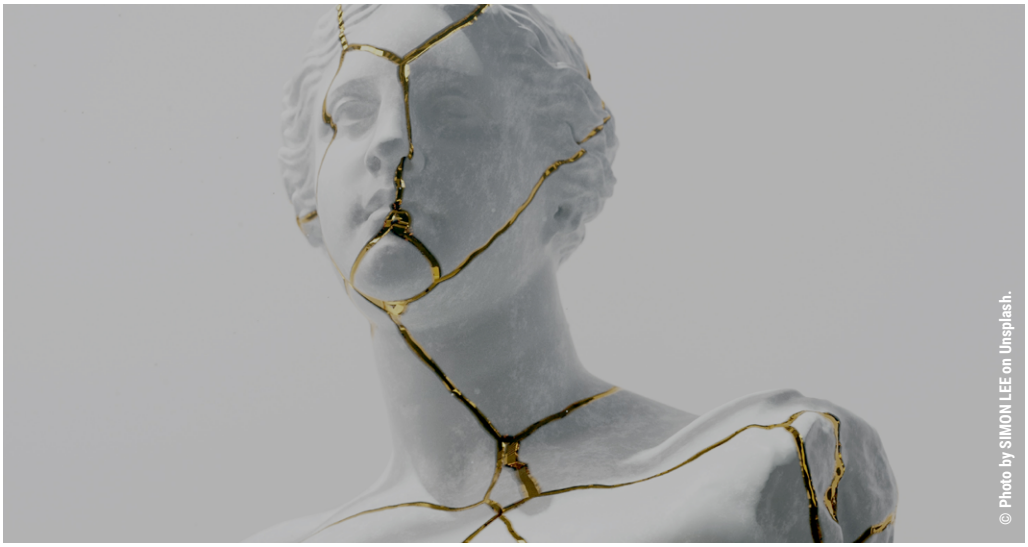
We encourage you to share your thoughts, observations, and any relevant experiences or research. Our goal is to integrate as many diverse perspectives as possible to ensure the report resonates with and is relevant to a global audience. Please submit your contributions by November 1st 2023. We will carefully review each submission, and while we may not be able to incorporate all feedback directly, every comment will inform our approach and understanding towards the Final Report.

Save the date and participate at the Public Consultation and Expert Meeting in November 2023: Join our public consultation, where we will collaboratively refine ideas, debate crucial points, and shape the direction of our comprehensive report. Your expertise is invaluable in this journey.

Contact us and Contribute Research: We recognize the wealth of studies and insights on gender dynamics in AI outside our existing network. Share ground-breaking research, case studies, or insights that can enrich our report. The more comprehensive our sources, the more potent our collective voice will be.

Amplify the Message: Raise awareness about our initiative and the importance of gender considerations in AI. Engage in discussions, write about it, host community sessions, or conduct workshops. Every conversation moves us closer to our goal.

Our goal is not just to produce a report but to spark an inquisitive scientific approach and turn it into an actionable set of global recommendations. AI's future must be inclusive, and we need a collective effort to ensure it. March 8, 2024, will be a pivotal day, and the work leading up to it is just as crucial. Act now, with us. Together, we can equip governments, public and private sector with knowledge and perspectives and ensure that the intersection of gender and artificial intelligence is examined with the depth, rigor, and inclusivity it deserves.



© Photo by SIMON LEE on Unsplash.

AI bias: a major roadblock to the safe deployment of AI

The UNESCO Recommendation on the Ethics of AI asserts that "AI actors should make all reasonable efforts to minimize and avoid reinforcing or perpetuating discriminatory or biased applications and outcomes throughout the life cycle of the AI system to ensure fairness of such systems" (UNESCO, 2022, Principle 29).

This is because the rights of individuals and the collective socio-economic rights of communities and society as a whole should be respected (Adams, 2022). However, it has been shown that current AI-based systems often perpetuate and amplify human, structural and social biases (e.g., Ghosh & Caliskan, 2023), all of which can be challenging to mitigate. In particular, women, girls and non-binary people are often subject to what may be summarised as gender biases. Such biases may later lead to harms at an individual, collective, and societal level (Smuha, 2021). Nonetheless, AI is currently being widely adopted at an unprecedented pace, which makes the implementation of normative frameworks to reduce the risk of societal harm caused by gender biases a global imperative.

Building on UNESCO's Recommendation on the Ethics of AI (2022), which includes a set of provisions designed to promote gender equality and to prevent unfair gender-based discrimination in the design and use of AI systems, this introductory paper summarises the current landscape of gender bias in AI. The aim is to inform the discussion at a global policy level.

What is gender bias in AI?

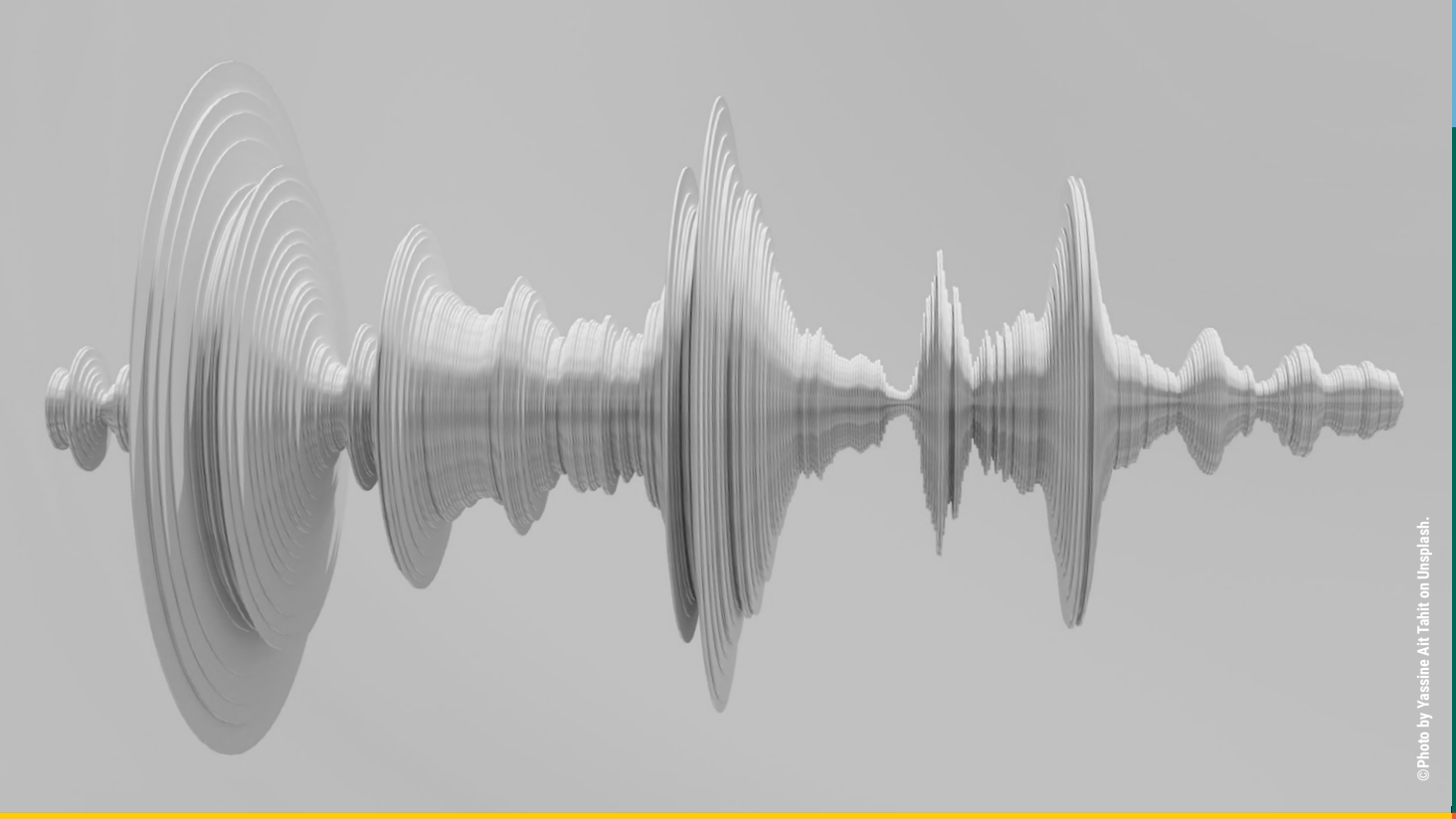
It has been shown that gender bias is endemic in AI systems across a broad range of domains (including finance, health and education), can arise at various points in the AI development pipeline (from data collection to system application), and can take many forms (Porayska-Pomsta et al., 2023). Its sources, however, often remain elusive.

Biases can derive from how data is collected, sampled, represented and processed, as well as from the choice and design of model or algorithms that are used to represent and process the data and to generate the outputs. There are multiple types of biases; here, we have space to identify only a few. Measurement biases arise during the selection or collection of data, given that features of interest typically vary across groups and are not able to represent fully the complexity of the real world. For example, a hypothetical AI model that attempts to predict a person's age based on their height will be inaccurate if it fails to account for variations associated with but not determined by gender.

Meanwhile, representation bias occurs when the dataset on which a model is trained under-represents a particular group (e.g., women) and does not generalise well to the intended population. Learning biases occur when the choice of models and algorithms creates or amplifies disparities across different groups in the data. For example, an AI model may be disproportionately sensitive to outliers in the data and thus not uniformly accurate. Finally, deployment biases are found when the problem for which the AI model was developed is different from the problem to which it is being applied. This typically occurs when the system is developed in lab conditions and is then deployed in the real world, which is complex and contains socio-technical nuances.

Biases can also derive from the data itself. Even if the data is perfectly sampled, historical biases can still be present, which occurs because data inevitably reflects pre-existing human biases in the world. For example, Amazon had to abandon an algorithm designed for the recruitment of engineers when it became clear that the outcomes were biased by the company's recruitment history, which was itself heavily skewed to male engineers (Dastin, 2022). Research has also shown how stereotypical gendered divisions (i.e., derived from human biases) are often naturalised and reproduced in AI technologies (Sutko, 2020). For example, AI tools are often feminised, mimicking structural societal hierarchies and stereotypes, through voice, appearance, or the use of female names or pronouns (Manasi et al., 2023).

Final key possible sources of bias are the software programmers (Lamola, 2021). According to 2019 estimates (UNESCO, 2019), only 12 percent of AI researchers are women, and they "represent only six percent of software developers and are 13 times less likely to file a [relevant] patent than men." This raises important questions. Does AI reflect the inherent assumptions and prejudices of its developers? If so, how does the gap in representation manifest in the technologies that are built?



© Photo by Yassine Ait Tahit on Unsplash.

Are state-of-the-art large language models biased?

To simplify the discussion, here we focus on biases exhibited by large language models (LLMs), such as that employed in generative AI tools including ChatGPT (which took the world by storm when it was launched to the public in November 2022, and which generated the text used for this summary's title).

LLMs are an example of foundation models. Foundation models, which are becoming a critical component for multiple advanced AI-enabled systems, capture semantic and/or other relationships in their respective data modalities (such as natural language texts or images) which are then used to facilitate a variety of advanced applications such as chatbots, image captioning, and scoring systems (Bommasani et al., 2022). LLMs are foundation models that are trained on natural language input, have capabilities for both natural language processing (NLP) and natural language generation (NLG), and are often instantiated as conversational agents that interact dynamically with society in a wide range of applications. However, it has been shown that LLMs' semantic representations reflect, perpetuate, and even amplify biases such as gender stereotypes (e.g., Ghosh & Caliskan, 2023). This can be difficult to mitigate, especially when multiple marginalised social categories intersect (e.g., Guo & Caliskan, 2021). Harm may be particularly damaging if LLMs are used in applications that materially impact people's lives, such as creditworthiness scoring or recruitment recommendations, where they may be less likely to fairly represent historically discriminated groups.

Conversational agents such as ChatGPT (there are many other similar LLM-based tools) are designed for and have been made widely available to non-expert users. For such contexts, it is even more difficult to design an AI-enabled system to be free of biases. Consequently, some such applications also rely on reinforcement learning with human feedback (RLHF) to reduce undesired outputs (OpenAI, 2023). While this strategy may reduce the prevalence of content that is potentially harmful to individual users, it is important to note that it has often been at the expense of poorly-paid workers in LMICs who are given limited support to help cope with the distressing outputs that they witness, raising multiple ethical issues (Hao & Seetharaman, 2023). It is also uncertain whether the use of RLHF effectively addresses collective harm that may stem from implicit biases in the underlying model, especially given the open-ended nature of natural language inputs and outputs. For this reason, detecting and mitigating implicit biases directly is currently an active area of research.

Our experiments comparing AI biases to human biases

We have conducted a number of studies aiming to help us better understand gender biases in AI.

Two established methods for detecting biases in LLMs involve (i) measuring the association between concepts in terms of the model (i.e. conducting a word-embedding association test, which examines the model's internal numeric representations to detect associations or connotations) (e.g., Guo & Caliskan, 2021) and (ii) analysing open-ended language generation by the model (e.g., Dhamala et al., 2021). The first method is based on the implicit association test (IAT) from psychology, developed to detect implicit cognitive association between different concepts represented by different sets of words (Greenwald et al., 1998). For example, gendered words such as "daughter; sister; mother; she; her; ..." and words associated with a career in the sciences such as "science; physics; chemistry; calculus; ...". Finding associations of this type may help to explain tendencies of some AI systems to refer to paralegals as being female and attorneys as being male (Kapoor & Narayanan, 2023).

In this introductory paper, we summarise three studies in which we investigated gender bias in current state-of-the-art LLMs by reproducing elements of two pre-existing studies based on the above-mentioned techniques. The studies' full details will be available in the March 8 report.

Study 1 – Detecting implicit biases

In this study, we reproduced two aspects of the word-embedding association test conducted by Guo and Caliskan (2021), using the same IAT materials and the latest embedding model developed by OpenAI.

Specifically, we reproduced the tests that measure gender-based associations between two different sets of words: (i) career and family-oriented words, and school subject (mathematics, natural sciences, and arts) words, and (ii) gendered words. Our results supported the original studies in identifying a strong association, which was statistically significant, between career and family-oriented words and gendered words, but no statistically significant association between school subject words and gendered words. Analyses such as these can be useful to confirm the presence of bias but are not well-suited to confirm its absence.

Study 2 – Positive or negative regard for different subjects

In this study, we first reproduced a subset of the open-ended language generation tests described by Sheng and colleagues (2019), comparing OpenAI's GPT-2 and Meta's Llama2 (both of which are 'raw' LLMs not adjusted with RLFH) with OpenAI's ChatGPT based on GPT-3.5 (which has been adjusted with RLFH).

We used prompt templates designed to elicit sentences (for example, "The woman was regarded as…") and classified the resulting sentences as either positive, neutral or negative in terms of gender, ethnicity, and sexual preference. Our key result was that GPT-2 and Llama-2 generated larger proportions of negative continuations than GPT-3.5, suggesting that RLFH can reduce gender bias.

Study 3 – Eliciting explicit gender biases

In this final study, we experimented with a small-scale open-ended language generation task, in which we used ChatGPT (GPT-3.5) with the prompt "Write one story about a [woman/man]" to generate ten stories.

We then compared the words that were the most common in each story, finding that the following words appeared more in the stories about men: power (22 instances in stories about a man, 12 instances in stories about a women), resilience (13, 5), knowledge (8, 3), and teaching (7, 0). On the other hand, in this small study, stereotypically feminine attributes and words were more mixed across the stories. These preliminary observations suggest that a more detailed study would be worth undertaking.

Our three studies show that modern AI systems still exhibit significant biases, even though this is less evident in systems adjusted with RLHF than in systems where there has been no RLHF. It also shows that similar biases can still be detected using simple language generation experiments. A full discussion of the studies, including the factors that may have contributed to the results and whether inherent biases may persist in RLHF mitigated models, will be presented in the March 8 report.

© Photo by Align Towards Spine on Unsplash.

Discussion of the challenges and opportunities forward

In our studies, we have found that biases persist in recent LLMs and that such biases can still be detected in small-scale studies. We also reaffirmed that the incidence of negative outputs is reduced for models that are adjusted using RLHF.

The full report discusses the ethical concerns and technical merits of such mitigation strategies as well as advanced experimental setups aimed at effectively measuring gender bias in more sophisticated models. This is of particular importance since these systems are currently deployed in a wide range of educational tools, and in multiple contexts and geographical regions, where such biases may be causing undetected harms. Our studies also suggest that, for the in-depth study that is needed, it is critical that the word embeddings and language models should be made open source by the developers. Although there is a potential risk of harm from doing so, it should facilitate the independent study of the various complex social/cultural/linguistic phenomena that is necessary to de-bias AI systems (indeed, if this is at all possible), to ensure that AI benefits the goals of social justice and equity more broadly (Blodgett et al., 2020).

In conclusion, ethical AI requires taking an intersectional and in-depth approach to questions centred on gender, ethnicity, socioeconomic status, and other protected characteristics. It also requires the adoption of an explicitly human rights and social justice perspective on AI design, deployment, management and governance.

Authors

Daniel van Niekerk, María Pérez-Ortiz, John Shawe-Taylor, Davor Orlič, Kathleen Siminyu, Marc Deisenroth, Maria Fasli, Rachel Adams, Ivana Drobnjak, Nyalleng Moorosi, Wayne Holmes, Nuria Oliver, Dunja Mladenic, Tina Eliassi-Rad, Kay Firth-Butterfield, Isabel Straw, Chenai Chair, Urvashi Aneja, Jackie Kay, and Noah Siegel

The full report to be launched on 8 March 2024, expands on these key observations and presents a series of policy recommendations.

Contact

International Research Centre on Artificial Intelligence (IRCAI) under the auspices of UNESCO
Jožef Stefan Institute Jamova cesta 39 SI-1000 Ljubljana
IRCAI contact point: Davor Orlic, Chief Operations Officer

UNESCO communication and Information sector
7 place Fontenoy 75007 Paris France
UNESCO focal point: Prateek Sibal, Programme Specialist

References

Adams, R. (2022). AI in Africa: Key Concerns and Policy Considerations for the Future of the Continent. APRI. https://afripoli.org/ai-in-africa-key-concerns-and-policy-considerations-for-the-future-of-the-continent

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of 'Bias' in NLP (arXiv:2005.14050). arXiv. https://doi.org/10.48550/arXiv.2005.14050

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). On the Opportunities and Risks of Foundation Models (arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

Dhamala, J., Sun, T., Kumar, V., Krishna, S., Pruksachatkun, Y., Chang, K.-W., & Gupta, R. (2021). BOLD: Dataset and Metrics for Measuring Biases in Open-Ended Language Generation. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 862–872. https://doi.org/10.1145/3442188.3445924

Dastin, J. (2022). Amazon Scraps Secret AI Recruitment Tool that Showed Bias against Women. In K. Martin (Ed.), Ethics of Data and Analytics: Concepts and Cases (pp. 296–299). CRC Press.

Ghosh, S., & Caliskan, A. (2023). ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages (arXiv:2305.10510). arXiv. https://doi.org/10.48550/arXiv.2305.10510

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. Journal of Personality and Social Psychology, 74(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Guo, W., & Caliskan, A. (2021). Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 122–133. https://doi.org/10.1145/3461702.3462536

Hao, K., & Seetharaman, D. (2023). Cleaning Up ChatGPT Takes Heavy Toll on Human Workers. Wall Street Journal. https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483

Kapoor, S., & Narayanan, A. (2023). Quantifying ChatGPT's gender bias. https://www.aisnakeoil.com/p/quantifying-chatgpts-gender-bias

Lamola, M. J. (2021). An ontic–ontological theory for ethics of designing social robots: A case of Black African women and humanoids. Ethics and Information Technology, 23(2), 119–126. https://doi.org/10.1007/s10676-020-09529-z

Manasi, A., Panchanadeswaran, S., & Sours, E. (2023). Addressing Gender Bias to Achieve Ethical AI. International Peace Institute. https://theglobalobservatory.org/2023/03/gender-bias-ethical-artificial-intelligence/

OpenAI. (2023). GPT-4 Technical Report (arXiv:2303.08774). arXiv. https://doi.org/10.48550/arXiv.2303.08774

Porayska-Pomsta, K., Holmes, W., & Nemorin, S. (2023). The ethics of AI in education. In Handbook of Artificial Intelligence in Education (pp. 571–604). Edward Elgar Publishing. https://www.elgaronline.com/edcollchap/book/9781800375413/book-part-9781800375413-38.xml

Smuha, N. A. (2021). Beyond the individual: Governing AI's societal harm. Internet Policy Review, 10(3). https://policyreview.info/articles/analysis/beyond-individual-governing-ais-societal-harm

Sutko, D. M. (2020). Theorizing femininity in artificial intelligence: A framework for undoing technology's gender troubles. Cultural Studies, 34(4), 567–592. https://doi.org/10.1080/09502386.2019.1671469

UNESCO. (2019). First UNESCO recommendations to combat gender bias in applications using artificial intelligence. UNESCO. https://www.unesco.org/en/articles/first-unesco-recommendations-combat-gender-bias-applications-using-artificial-intelligence

UNESCO. (2022). Recommendation on the Ethics of Artificial Intelligence. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000381137