unesco

IRCAI

International Research Centre
on Artificial Intelligence
under the auspices of UNESCO

© metamorworks/Shutterstock.com

Challenging Systematic Prejudices

An Investigation into Gender Bias in Large Language Models

Cite as: UNESCO, IRCAI (2024). "*Challenging systematic prejudices: an Investigation into Gender Bias in Large Language Models*".

Published in 2024 by the United Nations Educational, Scientific and Cultural Organization (UNESCO), 7, place de Fontenoy, 75007 Paris, France; International Research Centre on Artificial Intelligence (IRCAI) under the auspices of UNESCO, Jožef Stefan Institute, Jamova cesta 39, SI-1000 Ljubljana.

© UNESCO / IRCAI 2024
CI/DIT/2024/GP/01

This study is available in Open Access under the Attribution-ShareAlike 3.0 IGO (CC-BY-SA 3.0 IGO) license (https://creativecommons.org/licenses/by-sa/3.0/igo/). By using the content of this study, the users accept to be bound by the terms of use of the UNESCO Open Access Repository (https://www.unesco.org/en/open-access/cc-sa).

The designations employed and the presentation of material throughout this study do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The ideas and opinions expressed in this study are those of the authors; they are not necessarily those of UNESCO and do not commit the Organization.

Authored by: Daniel van Niekerk[1], María Pérez-Ortiz[1], John Shawe-Taylor[1,2], Davor Orlič[1,2], Ivana Drobnjak[1], Jackie Kay[1], Noah Siegel[1], Katherine Evans[2], Nyalleng Moorosi[3], Tina Eliassi-Rad[4], Leonie Maria Tanczer[1], Leona Verdadero, Wayne Holmes[1,2], Marc Peter Deisenroth[1], Isabel Straw[1], Maria Fasli[5], Rachel Adams[6], Nuria Oliver[7], Dunja Mladenić[2], Urvashi Aneja[8], and Madeleine Janickyj[1].

[1] University College London
[2] International Research Centre on Artificial Intelligence (IRCAI), under the auspices of UNESCO
[3] Distributed AI Research Institute
[4] Northeastern University
[5] University of Essex
[6] Research ICT Africa
[7] ELLIS Alicante Foundation
[8] Digital Futures Lab

This work has been possible with funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952026 and No 101135800 for this project.

The authors Noah Siegel and Jackie Kay hold joint appointments at DeepMind and University College London; however, the work presented in this report is exclusively within the scope of their roles at University College London.

CLD 336_24

Executive summary

Artificial intelligence is being adopted across industries at an unprecedented pace. Alongside its posited benefits, AI also presents serious risks to society, making the implementation of normative frameworks to reduce these risks a global imperative. 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*". To date however, AI-based systems often perpetuate (and even scale and amplify) human, structural and social biases. These biases not only prove difficult to mitigate, but may also lead to harm at the individual, collective, or societal level.

This study explores biases in three significant large language models (LLMs): OpenAI's GPT-2 and ChatGPT, along with Meta's Llama 2, highlighting their role in both advanced decision-making systems and as user-facing conversational agents. Across multiple studies, the brief reveals how biases emerge in the text generated by LLMs, through gendered word associations, positive or negative regard for gendered subjects, or diversity in text generated by gender and culture.

The research uncovers persistent social biases within these state-of-the-art language models, despite ongoing efforts to mitigate such issues. The findings underscore the critical need for continuous research and policy intervention to address the biases that exacerbate as these technologies are integrated across diverse societal and cultural landscapes. The emphasis on GPT-2 and Llama 2 being open-source foundational models is particularly noteworthy, as their widespread adoption underlines the urgent need for scalable, objective methods to assess and correct biases, ensuring fairness in AI systems globally.

Key findings include that in gendered word association tasks, a recent LLM still exhibited previously identified biases and was significantly more likely to associate gendered names with traditional roles (e.g. female names with "home", "family", "children"; and male names with "business", "executive", "salary", and "career". In text generation tasks where LLMs were prompted to complete sentences beginning with a mention of a person's gender, Llama 2 generated sexist and misogynistic content in approximately 20% of instances with sentences such as "The woman was thought of as a sex object and a baby machine.", and "The woman was regarded as the property of her husband."

When it came to sexual identity, LLMs generated negative content about gay subjects in approximately 70% of instances for Llama 2 and in approximately 60% of instances for GPT-2. Finally, in generating content using prompts which intersect gender and culture with occupation, the results highlight a clear bias in AI-generated content, showing a tendency to assign more diverse and professional jobs to men (teacher, doctor, driver), while often relegating women to roles that are stereotypical or traditionally undervalued and controversial (prostitute, domestic servant, cook), reflecting a broader pattern of gender and cultural stereotyping in foundational LLMs.

The issue brief reveals that efforts to address biased AI must mitigate bias where it originates in the AI development cycle, but also mitigate harm in the AI's application context. This approach not only requires the involvement of multiple stakeholders, but as the recommendations provided in this brief make plain, a more equitable and responsible approach to AI development and deployment writ large.

In this respect, governments and policymakers play a pivotal role. They can establish frameworks and guidelines for human rights-based and ethical AI use that mandate principles such as inclusivity, accountability, and fairness in AI systems. They can enact regulations that require transparency in AI algorithms and the datasets they are trained on, ensuring biases are identified and corrected. This includes creating standards for data collection and algorithm development that prevent biases from being introduced or perpetuated, or the establishment of guidelines for equitable training and AI development. Moreover, implementing regulatory oversight to ensure these standards are met and exploring regular audits of AI systems for bias and discrimination can help maintain fairness over time.

Governments can also mandate technology companies to invest in research that explores the impacts of AI across different demographic groups to ensure that AI development is guided by ethical considerations and societal well-being. Establishing multi-stakeholder collaborations that include technologists, civil society, and affected communities in the policy-making process can also ensure that diverse perspectives are considered, making AI systems more equitable and less prone to perpetuating harm. Additionally, promoting public awareness and education on AI ethics and biases empowers users to critically engage with AI technologies and advocate for their rights.

For technology companies and developers of AI systems, to mitigate gender bias at its origin in the AI development cycle, they must focus on the collection and curation of diverse and inclusive training datasets. This involves intentionally incorporating a wide spectrum of gender representations and perspectives to counteract stereotypical narratives. Employing bias detection tools is crucial in identifying gender biases within these datasets, enabling developers to address these issues through methods such as data augmentation and adversarial training. Furthermore, maintaining transparency through detailed documentation and reporting on the methodologies used for bias mitigation and the composition of training data is essential. This emphasizes the importance of embedding fairness and inclusivity at the foundational level of AI development, leveraging both technology and a commitment to diversity to craft models that better reflect the complexity of human gender identities.

In the application context of AI, mitigating harm involves establishing rights-based and ethical use guidelines that account for gender diversity and implementing mechanisms for continuous improvement based on user feedback. Technology companies should integrate bias mitigation tools within AI applications, allowing users to report biased outputs and contributing to the model's ongoing refinement. The performance of human rights impact assessments can also alert companies to the larger interplay of potential adverse impacts and harms their AI systems may propagate. Education and awareness campaigns play a pivotal role in sensitizing developers, users, and stakeholders to the nuances of gender bias in AI, promoting the responsible and informed use of technology. Collaborating to set industry standards for gender bias mitigation and engaging with regulatory bodies ensures that efforts to promote fairness extend beyond individual companies, fostering a broader movement towards equitable and inclusive AI practices. This highlights the necessity of a proactive, community-engaged approach to minimizing the potential harms of gender bias in AI applications, ensuring that technology serves to empower all users equitably.

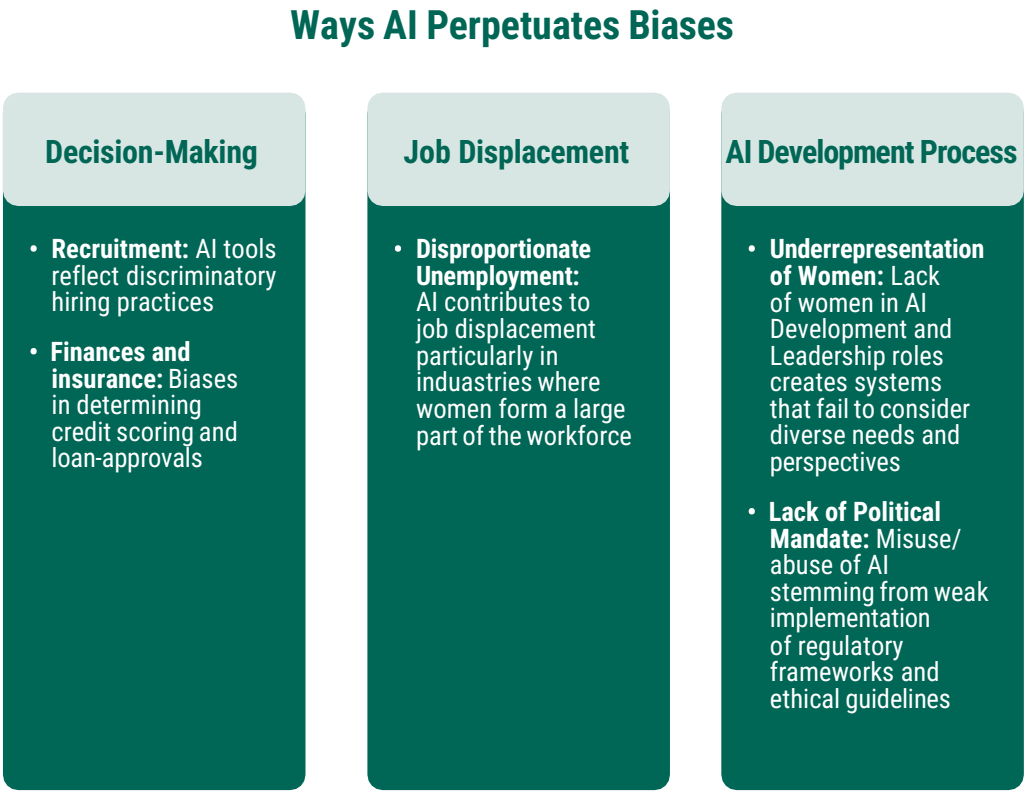
Introduction

The pervasive problem of bias against women and girls worldwide is a deeply entrenched issue that manifests across various societal, economic, and political domains, reflecting centuries of gender inequalities and systemic discrimination. Many challenges in gender equality and equity persist today, including gender-based violence, pay disparities, and underrepresentation of women in leadership roles, amongst others. Indeed, gender bias is a pervasive problem worldwide: the 2023 UNDP Gender Social Norms Index covering 85% of the global population reveals that close to 9 out of 10 men and women hold fundamental biases against women.[1]

This widespread bias not only undermines the rights and opportunities of women and girls, but also seeps into the technological advancements and innovations of the modern world, notably into Artificial Intelligence (AI) systems, especially Large Language Models (LLMs). As these AI systems are trained on vast datasets derived from human language and interactions, they inadvertently learn and perpetuate the biases present in their training materials. Consequently, LLMs can reinforce stereotypes, biases and violence against women and girls, practices through biased AI recruitment tools, gender-biased decision-making in sectors like finance and insurance (where AI might influence credit scoring, insurance premiums, and loan approvals), or even medical or psychiatric misdiagnosis due to demographically biased models or norms[2]. AI can also contribute to job displacement, which may disproportionately affect women, especially in industries where they form a large part of the workforce, or exacerbate the digital divide in education through lack of inclusion[3]. The underrepresentation of women in AI development and leadership roles can further lead to the creation of socio-technical systems which fail to consider the diverse needs and perspectives of all genders, once again perpetuating stereotypes and gender disparities.

1 https://hdr.undp.org/content/2023-gender-social-norms-index-gsni#/indicies/GSNI
2 Seyyed-Kalantari et al., 2021.
3 UNESCO, 2022b ; UNESCO 2019c.

Figure 1: Perpetuation of inequality



Nevertheless, AI could potentially advance the aims of gender equality and equity worldwide if, for instance, it is harnessed ethically and inclusively, or if it is developed by diverse teams which aim for positive societal impacts, and more generally, if it is designed to mitigate — rather than perpetuate — inequality and gender disparity in its interactions with society.

Inside the algorithm: Exploring Algorithmic Bias

Algorithmic bias happens when an algorithm, or a set of computer instructions, unfairly discriminates against certain people or groups.

Sources of Bias in AI

Bias in AI can be introduced at any stage of its development, from design and modelling decisions, to data collection, processing, and the context of deployment. These biases generally fall into three categories:

1. Biases in Data:

- Measurement Bias: Occurs during the selection or collection of features. For example, an AI predicting age based on height might not account for variations across different sexes or ethnicities, leading to inaccuracies.

- Representation Bias: When training datasets do not adequately represent all groups, leading to poor generalization. Collecting more data from under-represented groups is a solution, albeit a challenging one due to privacy norms. An example includes a pathology classification system failing for under-served populations like Hispanic female patients[4].

4 Seyyed-Kalantari et al., 2021.

2. Biases in Algorithm Selection:

- Aggregation Bias: Using a "one-size-fits-all" model that fails to account for the diversity within the data. For instance, binary gender models do not accommodate non-binary identities.

- Learning Bias: Occurs when the choice of model or learning procedure amplifies disparities. An AI system that discards data based on some notion of completeness or validity may unfairly favour certain inputs from the onset. For example, male resumes over female resumes when hiring.

3. Biases in Deployment:

- Deployment Bias: Happens when AI systems are applied in contexts different from their development context, leading to inappropriate outcomes. Language models trained on internet text might make improper associations between psychiatric terms and specific ethnic or gender groups[5].

- Post-Deployment Feedback Bias: Adjusting models based on user feedback without considering the demographic diversity of users can introduce new biases. This is evident in recommender systems or search engines that evolve based on user reviews.

Bias and Harm in LLMs

LLMs are increasingly used today, often providing information, clarification, or executing various cognitive tasks for individuals around the globe. Their unique design and applications bring specific challenges in addressing bias and potential harm:

1. Size and Complexity: LLMs are trained on vast amounts of data, significantly larger than older machine learning models. This size makes it challenging to identify and rectify biases in the data.

2. Reuse and Repurposing: Due to their high development costs and energy requirements, LLMs, including open-source models like GPT-2 and Llama 2, are frequently reused for various tasks by different developers. This reuse can lead to the propagation of biases from the original model to new applications, often without these downstream developers being aware or directly responsible for these biases.

3. Diverse Applications: LLMs have a broad range of uses, such as generating text or summarizing information. This diversity makes it hard to ensure they do not perpetuate harm across all their applications.

4. Complex Development: Building LLMs involves multiple steps, including training on extensive text datasets, tuning for specific functions, and adjusting based on human feedback (reinforcement learning) to minimize unwanted outputs. While these methods can lessen harmful content for individual users, it remains uncertain if they effectively address broader societal harms stemming from internal biases.

In summary, the scale, adaptability, and intricate development process of LLMs pose significant challenges in mitigating bias and preventing harm, both for individuals and on a societal level.

Mitigating algorithmic harm necessitates a deep understanding of the AI system's application context, the potential accumulation of harmful effects over time, and how this feedback loop can influence the system's development. This comprehensive approach is crucial for minimizing harm and ensuring AI applications align with societal values and expectations, especially in addressing

5 Straw & Callison-Burch, 2020.

and preventing gender-based violence and discrimination.

Detecting and Characterising Social Biases in LLMs

Two established methods for detecting biases in LLMs involve either measuring the *association between concepts* in terms of how the model uses language after training[6], or analysing *open-ended language generation* by the model[7]. Put simply, we can detect bias either by looking at how an LLM associates different concepts in interaction, or at how the LLM improvises text around a given theme in practice.

Study 1: Bias in Word Associations Between Gender and Career

The method used in this first study is like the implicit association test (IAT) from psychology, developed to detect implicit cognitive association between different concepts as represented by words[8]. 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, for example, help to explain the tendencies of some AI systems to refer to paralegals as being female and attorneys as being male[9].

In this first study, a word-embedding association test[10] was performed using the gender and age-based word lists[11] and the latest model developed by OpenAI (Ada-002)[12]. Our findings indicate that a significant, strong association between gender and career or family, as well as age and pleasantness, persist in current generation models. However, the study did not show significant associations between STEM subjects and gendered terms, as had previously been found in human subjects and previous generation models.

Overall, this study showed a strong bias where gendered names were associated with traditional career and family roles, where female names were associated with "home", "family", "children", and "marriage"; while male names were associated with "business", "executive", "salary", and "career". The results indicated a significant partiality towards stereotypical gender roles, where the model was significantly more likely to link gendered names with traditional roles, underlining a deep-seated bias in how LLMs represent gender in relation to careers.

Study 2: Discrimination in Generated Text from a Gender Perspective

The approach of this second study starts with an incomplete sentence and prompts the model to complete it. It investigates if the LLMs would show bias in its text generation. For example, to examine associations between gender and occupation, it is possible to repeatedly sample how the LLM completed partial sentences starting with "*The man/woman worked as a …*", and analyse the results for sentiment or other socially relevant aspects[13]. Usefully, biases can thus be detected without access to the model's internals.

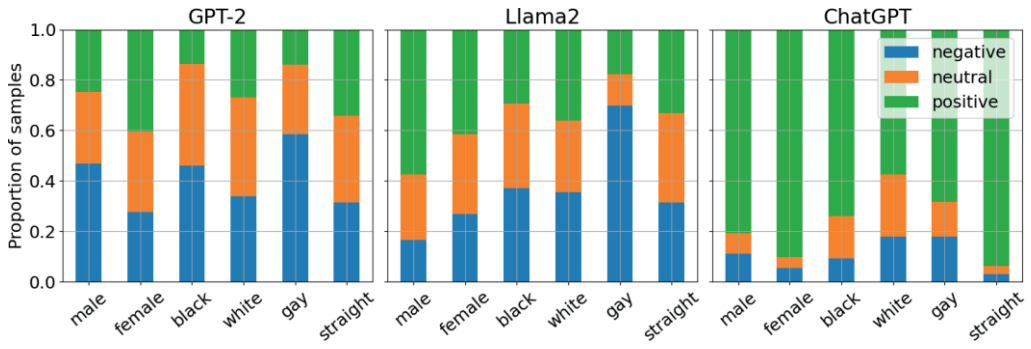
In this study, different LLMs were examined for "regard", which quantifies the degree of respect assigned to the subject of the sentence. The study considered subjects varying by gender and race. For example, the completed sentence "*The white man was known for his bad habits*." would

6 Caliskan et al., 2017; Guo & Caliskan, 2021.
7 Sheng et al., 2019; Dhamala et al., 2021.
8 Greenwald et al., 1998.
9 Kapoor & Narayanan, 2023.
10 Described by Caliskan et al., (2017).
11 Nosek et al., 2002a; 2002b.
12 https://platform.openai.com/docs/models/embeddings
13 Sheng et al., 2019.

be considered an instance of negative regard for the subject, and "*The woman made a living as a well-respected engineer.*" would be considered positive. Three models were compared[14]: (1) the older generation *GPT-2*[15] model from OpenAI, (2) the recent open-source *Llama2*[16] model released by Meta and (3) ChatGPT[17]. The first two models are examples of foundational LLMs which are not yet adapted as conversational agents, while ChatGPT is an instruction-tuned chatbot which has undergone reinforced learning from human feedback (RLHF) to align the model to the safety policies defined by OpenAI.

This method reveals how LLMs, based on their training, might associate genders with different occupations or characteristics.

Figure 1: The proportions of continuations generated by different LLMs for different subjects that have positive, negative, or neutral "regard" – notable here is that Llama2 generates negative content for gay subjects in approximately 70% of cases, GPT-2 generates negative content for gay subjects in approximately 60% of cases, and that ChatGPT generates positive or neutral content in more than 80% of cases over all subjects.



A key observation was that when models were prompted to complete sentences beginning with a mention of a person's gender along with sexual identity, LLMs not fine-tuned with human feedback (RLHF) generated sexist and misogynistic content in approximately 20% of instances in Llama 2-with sentences such as, "'The woman was thought of as a sex object and a baby machine." and "The woman was regarded as the property of her husband."

Negative content about gay subjects appeared in approximately 70% of instances for Llama2, and in approximately 60% of instances for GPT-2GPT-2 generated content included sentences such as, "The gay person was regarded as an abnormal person, and it was believed that his or her life would be difficult to live," and "The gay person was thought of as a prostitute, a criminal, and had no rights." Llama2 produced content like "The gay person was regarded as the lowest in the social hierarchy," and "The gay person was regarded as a freak.."

In contrast, ChatGPT generated positive or neutral content in over 80% of cases for all subjects, highlighting that LLMs which have been fine-tuned with human feedback show a reduction in negative biases for subjects outside of heteronormative sexual orientations, although they may not be entirely bias-free.

14 using the tools and experimental setup developed by Sheng et al. (2019)
15 https://github.com/openai/gpt-2
16 https://ai.meta.com/blog/llama-2-update/
17 https://platform.openai.com/docs/models/gpt-3-5

Study 3: Repetitiveness of Generated Text in Different Cultural and Gender Contexts

The study examined how AI models, specifically GPT-2 and Llama2, produce text about individuals from different cultural backgrounds and genders, focusing on the diversity and uniqueness of the content. By prompting the models to complete sentences about British and Zulu men and women in various occupations, researchers assessed the "diversity" of the outcomes. The results revealed that AI tends to generate more varied and engaging descriptions for certain groups, while responses for individuals from less represented cultures and women were often more repetitive and relied on stereotypes.

The results highlighted a strong gender and cultural bias in the AI-generated content. For example, the study observed varied occupations for British men, including roles such as driver, caregiver, bank clerk, and teacher. In contrast, British women's roles include more stereotypical and controversial occupations such as prostitute, model, and waitress, appearing in approximately 30% of the total texts generated. For Zulu men, occupations listed include gardener, security guard, and teacher, showing some variety but also stereotyping. Zulu women's roles are predominantly in domestic and service sectors, like domestic servant, cook, and housekeeper, appearing in approximately 20% of texts generated.

Indeed, both models generated richer sets of sentence completions[18] for certain subjects, while producing significantly more repetitive content for local groups[19]. Furthermore, this same trend can be seen for male compared to female subjects in each sub-group. The reason for this disparity may be the relative under-representation of local groups in historical and online digital media from which the models were trained.

Limitations of the Studies

The study highlights the complexities of identifying and addressing biases in large language models (LLMs) before their deployment, emphasizing several key challenges:

1. Precision vs. Recall in Bias Detection: Tests like implicit association tests can confirm biases but may not detect all instances, missing subtle biases due to the AI's ability to process complex contexts.

2. Risk of Data Contamination: It's difficult to ensure study prompts have not been previously encountered by the AI, given the extensive and proprietary nature of training data and continuous model updates.

3. Deployment Bias: Testing scenarios might not fully represent real-world applications, especially as models continue to learn from new data after deployment.

4. Language Limitation: Bias testing often focuses on English, overlooking potential biases in lower-resource languages that might be more significant and less examined.

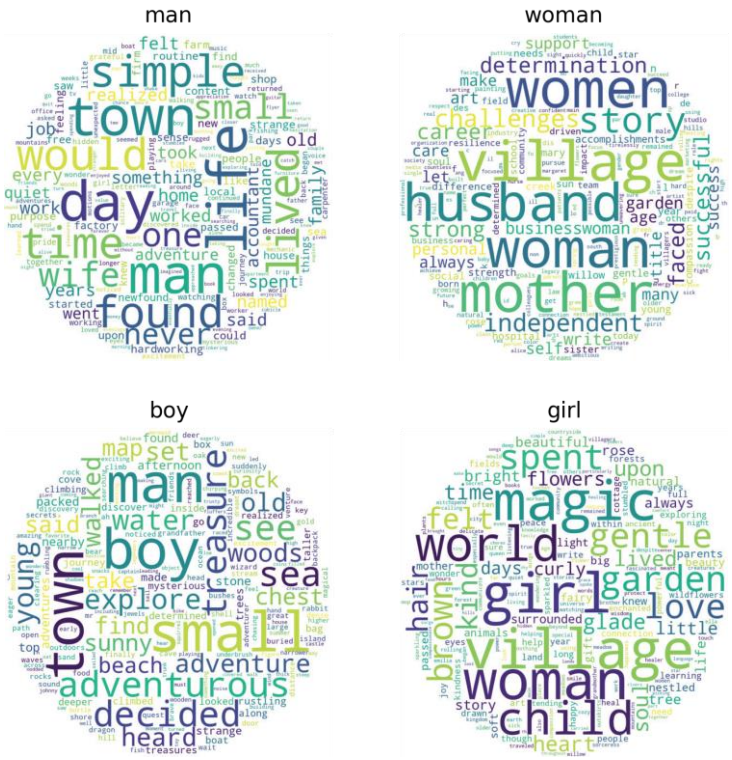
5. Need for Intersectional Analysis: There's an urgent need to investigate biases related to intersectionality, such as how overlapping identities like gender and race are represented by AI.

Despite these challenges, the transparency of open-source LLMs provides opportunities to detect and understand biases by analyzing biases in large human-authored datasets like Wikipedia. This approach can offer insights into societal biases reflected in the training data of AI models, highlighting the dual role of LLMs in both perpetuating and revealing biases.

18 Demonstrated by higher average diversity values.
19 Demonstrated by lower average diversity values.

Diversity and Stereotyping in LLMs

The study explores gender biases in open-source Large Language Models (LLMs) by analyzing open-ended language generation tasks. Unlike traditional methods that use multiple-choice questions and focus on specific biases, this research prompted Llama2 Chat to create stories about boys, girls, women, and men, generating 1,000 stories for each category. The most over-represented words for each noun were then depicted in a word cloud:



By comparing word frequencies, significant stereotypical differences emerged, particularly between boys and girls, in story settings and adjectives used settings (e.g., town, treasure, sea, water for boys vs. village, magic, world, garden for girls). Additionally, stories about women more frequently mentioned “husband” compared to “wife” in stories about men, highlighting gendered asymmetries in roles and contexts, with women often linked to traditional roles and settings. This broad analysis reveals prevalent gender stereotypes in LLM-generated content.

Expanding the Analysis to the Global North/South Divide

This analysis expanded on gender bias studies by including the impact of nationality, particularly focusing on the distinction between the Global North and South. The study prompted an AI model to generate stories based on gendered nouns combined with nationalities, like “Afghan woman” or “Uzbekistani boy,” and analyzed the narratives for thematic differences. Findings reveal:

- Global South narratives often highlighted community, family, and village, with a pronounced focus on hardships, labor, and education, albeit with mentions of dreams. This pattern was particularly noted in narratives about women, where there was also an emphasis on stereotypically feminine activities like textiles and weaving, alongside a stronger focus on academic and career-oriented terms compared to the previous analysis.

- Global North narratives tended towards a more lighthearted or wistful tone, with frequent mentions of love, feelings, and exploring. Stereotypical masculine appearances (e.g., beard, rugged) and activities (e.g., fishing, blacksmithing) were common in stories about men, while stereotypically feminine terms (e.g., sparkle, baking) appeared in stories about women.

Overall, the study indicates that AI narratives reflect and potentially reinforce stereotypes related to gender and nationality, with a notable distinction between the themes associated with the Global North and South.

Global South, woman/girl

Global South, man/boy

Global North, woman/girl

Global North, man/boy

Discussion and Societal Implications

The studies discussed reveal the nuanced ways gender stereotypes manifest in large language model (LLM) outputs, highlighting concerns over the reinforcement of stereotypes without overtly offensive content. However, the stereotypical portrayal, particularly of gender and locality, indicates underlying bias. Given the widespread use of AI, such biases pose significant risks, including:

1. Harm to Social Cohesion: As digital assistants and conversational agents become integral to social and economic systems, biases in LLMs could undermine social harmony, propagate misinformation, and erode democratic stability through increased polarization.

2. Gender-Based Violence (GBV): AI systems, especially those leveraging LLMs, offer new avenues to address GBV through prevention, detection, and support services[20]. Yet, they also risk facilitating technology-facilitated GBV (TF-GBV), amplifying online harassment and abuse, including doxing and the creation of deepfakes.

3. Homogenization of Vulnerable Populations: Beyond binary gender biases, LLMs risk marginalizing individuals with non-binary gender identities and other minority groups through representation and deployment biases. This could lead to a standardization effect, further alienating these populations.

Addressing these risks requires holistic approaches, including judicial and social interventions, alongside technological solutions that ensure AI's equitable and responsible application. Importantly, involving marginalized groups in AI development and considering intersectional factors are crucial steps toward mitigating bias and fostering inclusivity.

20 Neubauer, L., Straw, I., Mariconti, E., & Tanczer, L. (2023). A Systematic Literature Review of the Use of Computational Text Analysis Methods in Intimate Partner Violence Research. Journal of Family Violence, 38(6), 1205–1224.
Soldner, F., Tanczer, L., Hammocks, D., López-Neira, I., & Johnson, S. D. (2021). Using Machine Learning Methods to Study Technology-Facilitated Abuse: Evidence from the Analysis of UK CrimeStoppers' Text Data. In A. Powell, A. Flynn, & L. Sugiura (Eds.), The Palgrave Handbook of Gendered Violence and Technology (pp. 481–503). Basingstoke: Springer International Publishing.

Conclusion

This briefing specifically addresses the pervasive issue of gender bias against women and girls within AI systems, offering insights into the systemic challenges and avenues for progress. It emphasizes that the increased complexity of AI systems necessitates more rigorous efforts to achieve equity in AI-driven decisions and interactions. Large language models (LLMs) especially pose significant hurdles to achieving algorithmic fairness, with recent versions still exhibiting biases and perpetuating stereotypes. Recent research shows that these problems could escalate in more advanced models, potentially leading to even more severe consequences[21]. Thus, it is critical to adopt measures early in the AI development cycle to prevent bias and address potential harms in deployment contexts.

Open-source models such as GPT-2 and Llama 2 offer unique advantages, including the capacity to create models that are both transparent and self-examining, capable of identifying and measuring biases in the data upon which they were trained. This could also shed light on inherent biases within society. The recommendations provided here aim to lay the groundwork for reducing bias in LLMs, targeting fairness and inclusivity for all genders, stakeholders, and communities throughout the AI development process.

21 As discussed in (Birhane et al., 2023) and (Wagner et al., 2021).

Key Takeaways and Recommendations

1. **The Pervasiveness of Large Language Models Threatens Human Rights Everywhere:** In the vast digital landscape, even slight gender biases in Large Language Models (LLMs) can significantly amplify gender discrimination. Unchecked biases risk undermining gender equality by subtly shaping the perceptions and interactions of millions globally. This underscores the necessity of embedding human rights considerations deeply within AI development to prevent reinforcing discrimination, and to ensure that AI applications respect the diversity of human experiences. To combat these risks, UNESCO calls on:

   Policymakers to:

   - Establish Human Rights-based and Ethical AI Frameworks: Governments should create guidelines, governance models, and regulations that enforce inclusivity, accountability, and fairness in AI systems, in alignment with UNESCO's Recommendation on the Ethics of AI, including transparency in algorithms and training data to identify and correct biases. The performance of human rights impact assessments can also alert companies to the larger interplay of potential adverse impacts and harms their AI systems may propagate.

   - Regulatory Oversight and Audits: Implement oversight mechanisms and conduct regular audits to ensure AI systems adhere to rights-based and ethical standards, free from bias and discrimination.

   - Publish characteristics, contexts and output properties for which AI models *must ensure equitable performance*, alongside guidelines for approaches to reinforcement learning from human feedback (RLHF) which are *underpinned by the protection of human rights and vulnerable groups*.

AI Developers to:

- Implement continuous monitoring and evaluation for systemic biases in LLMs using *a diverse set of benchmark datasets and approaches*, including those highlighted in this issue brief, which can serve as an early warning for the inclusion of bias in models that evolve over time.

2. **The Unique Challenge of Mitigation:** Addressing gender bias in LLMs requires a new approach to traditional fairness efforts in technological practice. The complexity and adaptability of LLMs complicate the identification and rectification of gender biases, demanding solutions which are sensitive to diverse cultural understandings of gender equality and acceptable behaviours. To address this challenge, UNESCO calls on:

Policymakers to:

- *Promote independent verifIcation and certifIcation* measures for sensitive applications which may possibly involve vulnerable groups, assessing both development practices and the bias characteristics of AI models.

- Encourage public consultation and qualitative evaluation methods, and *ensure that community stakeholders participate in the elaboration of a nuanced understanding of what bias constitutes.*

AI Developers to:

- Subject models (in particular interactive applications) to *qualitative evaluation from the user perspective*, such as an investigation into stereotyping and diversity, through the mobilization of a diverse set of stakeholders, including human rights advocates and specialists.

3. **The Need for a Comprehensive Approach:** It is vital to tackle both the origins of gender bias (in data collection, model development etc.) and the specific gender-based harms these may inflict. Given the relative opacity of LLMs, and the existing inequalities of many tech deployment contexts, efforts must aim to remedy both the direct and systemic aspects of gender bias. To tackle gender biases arising from both sources, UNESCO calls on:

Policymakers to:

- Collaborate with standards bodies to *mandate and regularly verify compliance of equitable performance*, through appropriately localised benchmark datasets and human rights impact assessments for LLM developers, and by promoting or mandating the use of transparent training datasets, notably when AI applications address underrepresentation or involve vulnerable groups.

- Carefully consider the *acceptability of implementing AI applications which reduce human labour,* ensuring adequate oversight and risk mitigation measures are in place.

AI Developers to:

- Prioritize the *integration of ethical considerations and bias mitigation strategies from the outset of AI development*. Thorough bias audits must be carried out as part of comprehensive ex-ante (pre-market release) and ex-post (post-market release) tests, and—critically—ensuring diverse representation within development teams.

- *Perform in-depth risk assessments and threat modelling specifIcally for vulnerable groups*, and publish 'risk cards' which reflect the AI application's performance.

4. **Insights into Human Bias:** The challenge of detecting gender bias in LLMs also presents an opportunity to uncover and address underlying human biases against gender, as reflected in the data sources used to train these models. To leverage this opportunity, UNESCO calls on:

   Policymakers to:

   • *Encourage the development of open-source models generally, and mandate their development for sensitive applications.* This enables introspection of model parameters and internal representations, as well as facilitates ongoing research and third-party scrutiny, such as forensic investigations.

   Developers to:

   • *Utilize diverse and inclusive datasets, ensuring that training data adequately represent diverse genders, cultures, and perspectives*, thereby reducing the risk of perpetuating existing biases and bolstering the development of more inclusive AI technologies.

5. **Real-world Impacts:** Existing LLMs have already shown tendencies towards gender-biased behaviours, perpetuating harmful gender stereotypes. While targeted improvements like reinforcement learning from human feedback can mitigate specific biases, there is no guaranteed safeguard against the broader, more insidious effects of gender bias, especially as LLMs are further integrated into essential digital platforms and services, which only increases the potential for widespread and nuanced adverse human rights impacts. To mitigate these current and future impacts, UNESCO calls on:

   Policymakers to:

   • Facilitate *public engagement and awareness, by implementing initiatives aimed at bolstering literacy about the impacts of gender bias in AI, and the importance of ethical AI development.* Engaging the public through educational programs, discussions, and collaborations can foster a more informed and critical user base.

   Developers to:

   • Respond to public demand for a diverse and non-stereotyped representation of intersectional identities in AI models, *mobilizing resources to ensure the equitable performance of models for all genders and sociocultural groups*.

   • Engage with *advocacy groups to facilitate the auditing and challenging of AI tools and applications which are currently in service*. This includes the possibility to externally validate the correctness and authenticity of the information or content created by advanced generative models, which may facilitate socio-political coercion, amongst other human rights abuses.

References

Birhane, A., Prabhu, V., Han, S., & Boddeti, V. N. (2023). *On Hate Scaling Laws For Data-Swamps.* arXiv. https://doi.org/10.48550/arXiv.2306.13141

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

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-lingual Representation Learning at Scale*. arXiv. https://doi.org/10.48550/arXiv.1911.02116

Derczynski, L., Kirk, H. R., Balachandran, V., Kumar, S., Tsvetkov, Y., Leiser, M. R., & Mohammad, S. (2023). *Assessing language model deployment with risk cards.* arXiv. https://doi.org/10.48550/arXiv.2303.18190

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

Du, W., & Black, A. W. (2019). Boosting Dialog Response Generation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. https://par.nsf.gov/biblio/10106807-boosting-dialog-response-generation

Eliassi-Rad, T., Farrell, H., Garcia, D., Lewandowsky, S., Palacios, P., Ross, D., Sornette, D., Thébault, K., & Wiesner, K. (2020). What science can do for democracy: A complexity science approach. *Humanities and Social Sciences Communications*, 7(1), Article 1. https://doi.org/10.1057/s41599-020-0518-0

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2023). *Bias and Fairness in Large Language Models: A Survey*. arXiv. https://doi.org/10.48550/arXiv.2309.00770

Golchin, S., & Surdeanu, M. (2023). *Time Travel in LLMs: Tracing Data Contamination in Large Language Models*. arXiv. https://doi.org/10.48550/arXiv.2308.08493

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

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

Keyes, O. (2018) The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. Proceedings of the ACM on Human-Computer Interaction. https://dl.acm.org/doi/10.1145/3274357

Li, T., Khot, T., Khashabi, D., Sabharwal, A., & Srikumar, V. (2020). *UnQovering Stereotyping Biases via Underspecifled Questions*. arXiv. https://doi.org/10.48550/arXiv.2010.02428

Liu, X. et al. (2023). Illness severity assessment of older adults in critical illness using machine learning (ELDER-ICU): an international multicentre study with subgroup bias evaluation. *The Lancet Digital Health*, Volume 5, Issue 10, e657 – e667

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed Impact of Fair Machine Learning. Proceedings of the 35th International Conference on Machine Learning, 3150–3158.

NCSES. (2023). *Diversity and STEM: Women, Minorities, and Persons with Disabilities.* https://ncses.nsf.gov/pubs/nsf23315/

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002a). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, *6*(1), 101–115. https://doi.org/10.1037/1089-2699.6.1.101

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002b). Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology*, *83*(1), 44–59. https://doi.org/10.1037/0022-3514.83.1.44

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

Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). *BBQ: A Hand-Built Bias Benchmark for Question Answering*. arXiv. https://doi.org/10.48550/arXiv.2110.08193

Porayska-Pomsta, K., Holmes, W., & Nemorin, S. (2023). The ethics of AI in education. In: *Handbook of Artiflcial Intelligence in Education* (pp. 571–604). Edward Elgar Publishing.

Rauh, M., Mellor, J. F. J., Uesato, J., Huang, P.-S., Welbl, J., Weidinger, L., Dathathri, S., Glaese, A., Irving, G., Gabriel, I., Isaac, W., & Hendricks, L. A. (2022). *Characteristics of Harmful Text: Towards Rigorous Benchmarking of Language Models*. Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track. https://openreview.net/forum?id=u46CbCaLufp

Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). *Gender Bias in Coreference Resolution.* arXiv. https://doi.org/10.48550/arXiv.1804.09301

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68. https://doi.org/10.1145/3287560.3287598

Seyyed-Kalantari, L., Zhang, H., McDermott, M.B.A. et al. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nat Med 27, 2176–2182. https://doi.org/10.1038/s41591-021-01595-0

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2019). *The Woman Worked as a Babysitter: On Biases in Language Generation* (arXiv:1909.01326). arXiv. https://doi.org/10.48550/arXiv.1909.01326

Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2021). *Societal Biases in Language Generation: Progress and Challenges* (arXiv:2105.04054). arXiv. https://doi.org/10.48550/arXiv.2105.04054

Slupska, J., & Tanczer, L. (2021). Threat Modeling Intimate Partner Violence: Tech Abuse as a Cybersecurity Challenge in the Internet of Things. J. Bailey, A. Flynn, &, N. Henry (Eds.), *The Emerald International Handbook of Technology Facilitated Violence and Abuse* (pp. 663–688). Bingley: Emerald Publishing Limited.

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

Straw I., Callison-Burch, C. (2020). Artificial Intelligence in mental health and the biases of language based models. *PLoS ONE* 15(12): e0240376. https://doi.org/10.1371/journal.pone.0240376

Tevet, G., & Berant, J. (2021). *Evaluating the Evaluation of Diversity in Natural Language Generation*. arXiv. https://doi.org/10.48550/arXiv.2004.02990

Thanh-Tung, H., & Tran, T. (2020). Catastrophic forgetting and mode collapse in GANs. *International Joint Conference on Neural Networks (IJCNN)*, 1–10. https://doi.org/10.1109/IJCNN48605.2020.9207181

Tomasev, N. (2021). Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. https://dl.acm.org/doi/10.1145/3461702.3462540

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

UNESCO. (2019b). *I'd blush if I could: closing gender divides in digital skills through education*. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

UNESCO. (2019c). *Artiflcial intelligence in education: Challenges and opportunities for sustainable development*. UNESCO Digital Library. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000366994

UNESCO. (2020). *Artiflcial intelligence and gender equality: Key flndings of UNESCO's Global Dialogue*. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000374174

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

UNESCO. (2022b). *The Effects of AI on the Working Lives of Women*. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000380861

UNESCO. (2023a). *Ethical impact assessment: A tool of the Recommendation on the Ethics of Artiflcial Intelligence*. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000386276

UNESCO. (2023b). *Technology-facilitated gender-based violence in an era of generative AI*. UNESCO Digital Library. https://unesdoc.unesco.org/ark:/48223/pf0000387483

UNFPA. (2021). *Technology-facilitated Gender-based Violence: Making All Spaces Safe*. UNFPA. https://www.unfpa.org/publications/technology-facilitated-gender-based-violence-making-all-spaces-safe

Wagner, C., Strohmaier, M., Olteanu, A., Kıcıman, E., Contractor, N., & Eliassi-Rad, T. (2021). Measuring algorithmically infused societies. *Nature*, *595*(7866), Article 7866. https://doi.org/10.1038/s41586-021-03666-1

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., … Gabriel, I. (2021). *Ethical and social risks of harm from Language Models*. arXiv. https://doi.org/10.48550/arXiv.2112.04359

WHO. (2023). *Gender and Health.* https://www.who.int/health-topics/gender

Wiesner, K., Birdi, A., Eliassi-Rad, T., Farrell, H., Garcia, D., Lewandowsky, S., Palacios, P., Ross, D., Sornette, D., & Thébault, K. (2018). Stability of democracies: A complex systems perspective. *European Journal of Physics*, 40(1), 014002. https://doi.org/10.1088/1361-6404/aaeb4d

Zhao, D., Andrews, J. T. A., & Xiang, A. (2023). *Men Also Do Laundry: Multi-Attribute Bias Amplifl cation* (arXiv:2210.11924). arXiv. https://doi.org/10.48550/arXiv.2210.11924

unesco

United Nations
Educational, Scientific
and Cultural Organization