ISSN 2456-4842

Pahal Horizon
AN INTERNATIONAL RESEARCH JOURNAL
Homepage: www.pahalhorizon.com

Bias in AI: A Review of Current Research

Kamya Mahajan*
* Department of Psychology, University of San Francisco, San Francisco

Abstract- This research review explores the issue of racial and gender biases embedded within Artificial Intelligence (AI) systems. As AI technology is rapidly integrating into our daily lives, a variety of applications such as content curation, facial recognition, and hiring decisions reveal significant biases, adversely affecting marginalized communities. AI systems, powered by machine learning, often reflect and amplify existing societal prejudices due to imbalanced training data and non-inclusive algorithm designs. Studies show facial recognition systems frequently misidentify women and individuals with darker skin tones, while language models perpetuate gender stereotypes and traditional gender roles. This research review highlights long-standing issues of gender and racial bias, which persist across social, economic, and political domains, and are now reflected in AI technologies. This paper aims to analyze the root causes of these biases, their impacts on marginalized communities, and propose potential solutions. By understanding the scope and nature of these biases, the goal is to develop more ethical, fair, and inclusive AI systems that uphold principle of accountability, thus contributing to the broader effort towards gender equality and racial equity.

Index Terms- Artificial Intelligence (AI), Gender Bias, Racial Bias, Equity, Equality

I. Introduction

the rapidly developing field of Artificial Intelligence (AI) involves developing computer systems that are able to carry out tasks like learning, problem-solving, and decision-making that normally need human intelligence. Artificial Intelligence (AI) systems aim to emulate, if not exceed, human cognitive capacities in order to build computers that are capable of independent thought, perception, and behaviour. Nowadays, AI systems are being used more and more in a variety of applications. However, research has shown that these systems have notable racial and gender biases, which can have a detrimental effect on these marginalised communities (Lazaro, 2022).

The speed with which AI has been adopted has changed the way we use technology on a daily basis. These days, AI-powered recommendation algorithms, automated decision-making systems, and virtual assistants are common in offices, homes, and public areas (Manasi et al., 2022). While AI has increased our efficiency, data and algorithms that support AI systems have the potential to reinforce and magnify societal biases (Lazaro, 2022). Machine learning is the foundation of artificial intelligence systems. It involves training algorithms on massive datasets to find patterns and provide predictions. However, AI models may display biases along gender, racial, and other demographic lines if the training data is unbalanced or if the algorithms are not created with diversity and inclusion in mind. These prejudices may result in unfair consequences, such as the ranking of female job applications lower than that of male applicants with comparable qualifications or a higher rate of incorrect identifications of people of colour by facial recognition software.

The goal of this research review is to present a thorough analysis of the current situation with regard to racial and gender bias in AI, looking at the root causes, the

effects on marginalised people, and possible solutions to these pressing problems. By comprehending the kind and scope of these prejudices, it aims to create awareness to create more moral and inclusive AI systems that respect the rights of people, justice, and accountability.

II. Gender Bias in AI

Researchers have been exploring whether there are gender biases embedded within artificial intelligence (AI) systems in an effort to uncover troubling patterns that can undermine efforts towards greater gender equality. One such research study was conducted by UNESCO, titled "Bias Against Women and Girls in Large Language Models (2024)." This study aimed to provide a comprehensive examination of the prevalence of gender biases in some of the most widely used artificial intelligence (AI) systems. It was commissioned by the UNESCO Chair in AI at University College London, the study analysed the content generated by popular large language models (LLMs) such as GPT-3.5, GPT-2, and Llama 2 to uncover the extent and nature of gender stereotyping embedded within these AI tools. The researchers used a multi-pronged approach to assess the gender biases in the LLMs. This included measuring the diversity of content in AI-generated texts focused on individuals across a spectrum of genders, sexualities, and cultural backgrounds (UNESCO, 2024). The study also analysed the semantic associations between gender-specific names and various words and occupations in the generated text. Additionally, the researchers examined the prevalence of gender-based stereotypes and the assignment of traditional gender roles in the AI-produced narratives (UNESCO, 2024).

The findings of the study clearly demonstrated the pervasive gender biases present in the examined LLMs. The open-source models, such as Llama 2, exhibited the most significant biases, with a tendency to assign more diverse and high-status jobs to men while relegating women to traditionally undervalued or stigmatised roles (UNESCO, 2024). The language used to describe men and women also reflected stark differences, with stories about men dominated by words like "treasure," "adventurous," and "decided," while those about women emphasised domestic themes like "garden," "love," and "husband (UNESCO, 2024)."

These results underscore the urgent need to address the systemic gender biases embedded within the AI systems that are increasingly shaping our digital landscape (UNESCO, 2024). As these LLMs become more present in various applications, from content curation to decision-making, the perpetuation of gender stereotypes can have far-reaching consequences, potentially exacerbating existing inequalities and undermining efforts towards gender equality. The study's findings highlight the critical importance of developing more inclusive and ethical AI frameworks that prioritise fairness, diversity, and human rights.

Similarly, another study titled "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" provides a comprehensive examination of the gender biases present in leading facial analysis systems. Conducted by researchers at the MIT Media Lab and Microsoft Research, this study employed a rigorous, intersectional approach to evaluate the performance of three major commercial gender classification APIs - Microsoft, IBM, and Face++ (Buolamwini & Timnit Gebru, 2018). The purpose of this study was to assess the accuracy of these facial analysis systems across different demographic subgroups, with a particular focus on the intersection of gender and skin type. The researchers utilized the Fitzpatrick skin type scale, a dermatologist-approved classification system, to characterize the skin tones of the subjects in their newly created Pilot Parliaments Benchmark (PPB) dataset (Buolamwini & Timnit Gebru, 2018). They then evaluated the gender classification accuracy of the commercial APIs on this diverse dataset.

The results were deeply concerning, revealing significant disparities in the performance of these facial analysis systems (Buolamwini & Timnit Gebru, 2018). The researchers found that the error rates for classifying the gender of darker-skinned women were as high as 34.7%, while the error rates for lighter-skinned men were less than 1%. This stark contrast highlights the systemic biases embedded within these AI-powered technologies, which tend to perform best on the demographic groups that are overrepresented in their training data - in this case, lighter-skinned individuals (Buolamwini & Timnit Gebru, 2018).

These findings underscore the urgent need to address the lack of diversity and inclusivity in the development of facial analysis systems. As these technologies become increasingly ubiquitous in high-stakes applications, such as law enforcement and hiring decisions, the perpetuation of gender and racial biases can have

far-reaching consequences, exacerbating existing inequalities and undermining efforts towards greater social justice. The study's authors emphasise the critical importance of comprehensive algorithmic auditing and the inclusion of diverse perspectives in the design and deployment of these AI systems.

Another study titled "Gender Bias in Large Language Models: The Case of Persona-Chatbots" provides a comprehensive examination of the gender biases present in large language models (LLMs) used for persona-based chatbots (O'Connor et al., 2023). Conducted by researchers at the University of Cambridge and the University of Trento, this study employed a novel approach to assess the gender stereotyping in AI-generated narratives. The purpose of this study was to investigate the extent to which LLMs, when used to create persona-based chatbots, perpetuate gender stereotypes and assign traditional gender roles to the generated characters. The researchers developed a framework for analysing the semantic associations between gender-specific names and various words and occupations in the chatbots' responses. They also examined the prevalence of gender-based stereotypes and the assignment of traditional gender roles in the AI-produced narratives (O'Connor et al., 2023).

The results of the study were consistent with the findings of the UNESCO study, demonstrating the pervasive gender biases present in the examined LLMs (O'Connor et al., 2023). The researchers found that the chatbots exhibited a tendency to assign more diverse and high-status jobs to male characters while relegating female characters to traditionally undervalued or stigmatized roles. The language used to describe male and female characters also reflected stark differences, with narratives about male characters dominated by words like "leader," "ambitious," and "confident," while those about female characters emphasized domestic themes like "caring," "emotional," and "nurturing. (O'Connor et al., 2023)"

These findings too underscore the urgent need to address the systemic gender biases embedded within the AI systems that are increasingly shaping our digital landscape. As persona-based chatbots become more prevalent in various applications, from customer service to mental health support, the perpetuation of gender stereotypes can have far-reaching consequences, potentially reinforcing harmful societal norms and undermining efforts towards gender equality. The study's authors emphasize the critical importance of developing more inclusive and ethical AI frameworks that prioritize fairness, diversity, and human rights.

III. Racial Bias in AI

In addition to the pervasive gender biases uncovered in AI systems, researchers have also documented the prevalence of racial biases embedded within these technologies. A case study conducted for the World Health Organization (WHO) provides an examination of the issue of discrimination and racial bias in AI-powered applications (Minssen et al., 2021). The purpose of this study was to investigate the extent and nature of racial biases present in various AI technologies, with a particular focus on the implications for public health and healthcare delivery (Minssen et al., 2021). The researchers employed a multi-faceted approach, analysing the performance of AI systems across different domains, including facial recognition, natural language processing, and automated decision-making.

The findings of the study were concerning, revealing the systemic nature of racial biases in AI. Minssen et al. (2021) found that facial recognition algorithms exhibited significantly higher error rates when identifying individuals with darker skin tones, potentially leading to discriminatory outcomes in security and law enforcement applications. Similarly, natural language processing models were shown to exhibit biases in their language associations, with words related to certain racial groups being linked to more negative connotations (Minssen et al., 2021). These results underscore the urgent need to address the root causes of racial biases in AI development and deployment. The study highlights how the lack of diversity and representation within the AI workforce, as well as the use of training data that reflects societal prejudices, can perpetuate and amplify existing inequities. Furthermore, the researchers emphasise that the problem of racial bias in AI is not limited to technical factors, but is also deeply intertwined with broader societal and institutional structures that perpetuate discrimination.

To mitigate these issues, the study calls for a multifaceted approach that involves diversifying the AI development community, implementing rigorous bias testing and auditing procedures, and collaborating with affected communities to better understand the real-world

impacts of these biases (Minssen et al., 2021). By addressing the systemic nature of racial biases in AI, we can work towards the development of more equitable and inclusive technologies that uphold the principles of human rights and social justice.

Similarly, another study titled "Racial Bias in Deep Learning Models for the Classification of Chest X-Rays" provides critical insights into the prevalence of racial biases in AI-powered medical imaging analysis (Sham et al., 2023). Conducted by researchers at the University of Virginia, this study examined the performance of deep learning models in predicting a patient's race from chest X-ray images. The purpose of this study was to investigate the extent to which AI systems can accurately infer a patient's race from medical images that do not contain any explicit racial information. The researchers developed deep learning models trained on a large dataset of chest X-rays and evaluated their ability to predict the self-reported race of the patients (Sham et al., 2023).

The results of the study were very concerning, revealing the troubling presence of racial biases in these AI-powered medical imaging systems. The researchers found that the deep learning models were able to accurately predict a patient's race with high accuracy, even when the X-ray images did not contain any visible racial characteristics (Sham et al., 2023). This suggests that the models were learning to identify subtle, and potentially unintended, racial cues within the medical images. These findings underscore the urgent need to address the issue of racial bias in AI-powered medical technologies. As these systems become more widely adopted in healthcare settings, the ability to infer a patient's race from medical images could lead to discriminatory practices, such as differential treatment or resource allocation based on perceived racial characteristics. Moreover, the perpetuation of these biases in medical AI could exacerbate existing health disparities and undermine efforts to achieve equitable healthcare outcomes (Sham et al., 2023).

The researchers emphasize the critical importance of developing more inclusive and transparent AI frameworks for medical imaging analysis. This requires a concerted effort to diversify the AI development community, ensure the use of representative and unbiased training data, and implement rigorous bias testing and auditing procedures. By addressing the systemic nature of racial biases in medical AI, we can work towards the creation of more equitable and trustworthy technologies that uphold the principles of patient-centred care and human rights.

IV. Discussion

The body of research reviewed here highlights the significant and pervasive nature of gender and racial biases embedded within AI systems. These biases are not just technical issues but reflect broader societal prejudices that have found their way into the algorithms and datasets used to train these AI technologies (UNESCO, 2024) . The studies discussed, ranging from the UNESCO study on gender biases in large language models (LLMs) to the intersectional analysis of facial recognition systems and the assessment of racial biases in medical imaging, collectively underscore the multifaceted challenges posed by biased AI.

The UNESCO study on LLMs, such as GPT-3.5, GPT-2, and Llama 2, demonstrates that gender biases manifest in the content generated by these systems. The tendency to assign more diverse and high-status jobs to men while relegating women to traditional and undervalued roles points to an entrenched bias that could perpetuate gender inequality (UNESCO, 2024).. Moreover, the stark differences in language used to describe men and women—where men are associated with terms like "treasure" and "adventurous," and women with "garden" and "husband"—reveal a deep-seated stereotyping that could influence public perceptions and decisions influenced by AI-generated content.

Similarly, the "Gender Shades" study illustrates the intersectional nature of biases in commercial facial analysis systems (Buolamwini & Timnit Gebru, 2018). The high error rates for classifying the gender of darker-skinned women compared to lighter-skinned men highlight the compounding effect of racial and gender biases. These disparities not only undermine the accuracy of these systems but also pose significant risks in applications like law enforcement and hiring, where biased outcomes can have severe consequences.

The case study conducted for the WHO on racial bias in AI applications further emphasizes the systemic nature of these issues. The findings that facial recognition algorithms perform poorly on darker-skinned individuals and that NLP models exhibit biased language associations reinforce the need for a comprehensive approach to mitigating AI biases (Minssen et al., 2021). This study, along with the research on deep learning models for chest

X-ray classification, reveals that even in domains like healthcare, where AI holds great promise, the presence of biases can lead to discriminatory practices and exacerbate existing inequalities.

The pervasive nature of these biases calls for urgent interventions. Addressing the lack of diversity and representation within the AI development community is critical. A diverse workforce is more likely to identify and mitigate biases that may be overlooked by homogenous teams. Additionally, rigorous bias testing and algorithmic auditing are essential to identify and address biases before AI systems are deployed. Engaging with affected communities to understand the real-world impacts of biased AI is also crucial for developing technologies that are fair and inclusive.

V. CONCLUSION

The studies reviewed provide evidence of the urgent need to address gender and racial biases in AI systems. These biases are not merely technical flaws but reflections of broader societal inequities that have been encoded into AI technologies. The findings highlight the significant risks posed by biased AI, from perpetuating gender stereotypes and reinforcing racial discrimination to undermining efforts towards gender equality and social justice.

To mitigate these risks, a comprehensive approach is needed that involves diversifying the AI workforce, implementing rigorous bias testing and auditing procedures, and engaging with affected communities. Developing inclusive and ethical AI frameworks that prioritize fairness, diversity, and human rights is critical. By addressing the systemic nature of biases in AI, we can work towards the creation of technologies that not only advance innovation but also uphold the principles of equity and social justice. The future of AI should be such that technology serves all of humanity without perpetuating historical and societal prejudices.

VI. ACKNOWLEDGEMENTS.

I WOULD LIKE TO EXTEND MY HEARTFELT THANKS TO PAHAL HORIZON FOR OFFERING ME THIS INTERNSHIP OPPORTUNITY, ALL THE TEACHERS FOR THEIR INVALUABLE GUIDANCE, AND MY PARENTS FOR THEIR UNWAVERING SUPPORT AND ENCOURAGEMENT.

REFERENCES

[1] Arcilla, A. O., Espallardo, A. K. V., Gomez, C. a. J., Viado, E. M. P., Ladion, V. J. T., Naanep, R. a. T., Pascual, A. R. L., Artificio, E. B., & Tubola, O. D. (2023). Ethics in AI Governance: Comparative analysis, implication, and policy recommendations for the Philippines. International Computer Science and Engineering Conference (ICSEC). https://doi.org/10.1109/icsec59635.2023.10329756

[2] Buolamwini, J., & Timnit Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, (81), 1–15.

[3] García-Ull, Francisco-José; Melero-Lázaro, Mónica (2023). "Gender stereotypes in AI-generated images".Profesional de la información, v. 32, n. 5, e320505.https://doi.org/10.3145/epi.2023.sep.05 (PDF) Gender stereotypes in AI-generated images. Available from: https://www.researchgate.net/publication/373371887_Gender_stereotypes_in_AI-generated_images

[4] GO, A. (2023, March 17). Addressing gender bias to achieve ethical AI. IPI Global Observatory. https://theglobalobservatory.org/2023/03/gender-bias-ethical-artificial-intelligence/

[5] Lazaro, G. (2022, December 24). Understanding gender and racial bias in AI — Harvard ALI Social Impact Review. Harvard ALI Social Impact Review. https://www.sir.advancedleadership.harvard.edu/articles/understanding-gender-and-racial-bias-in-ai

[6] Manasi, A., Panchanadeswaran, S., Sours, E., & Lee, S. J. (2022). Mirroring the bias: gender and artificial intelligence. Gender, Technology and Development, 26(3), 295–305. https://doi.org/10.1080/09718524.2022.2128254

[7] Minssen, Timo & Gerke, Sara & Corrales Compagnucci, Marcelo. (2021). Discrimination and racial bias in AI technology: A case study for the WHO.

[8] O'Connor, S., Liu, H. Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. AI & Soc (2023). https://doi.org/10.1007/s00146-023-01675-4

[9] Sham, A.H., Aktas, K., Rizhinashvili, D. et al. Ethical AI in facial expression analysis: racial bias. SIViP 17, 399–406 (2023). https://doi.org/10.1007/s11760-022-02246-8

[10] UNESCO & International Research Centre on Artificial Intelligence. (2024). Challenging systematic prejudices: an investigation into bias against women and girls in large language models (Vol. 1). https://unesdoc.unesco.org/ark:/48223/pf0000388971?posInSet=1&queryId=N-EXPLORE-a85080c5-1aaa-4e57-bde6-b1f8cd29221a

AUTHORS

First Author – Kamya Mahajan, Student, University of San Francisco, kmahajan2@dons.usfca.ed