Gender Bias in Generative Artificial Intelligence: A Literature Review on Perspectives and Implications

Marcelo Pimentel[1] and José Carlos Veliz Palomino[2]
[1]Universidad de Lima, Peru
[2]CENTRUM Católica Graduate Business School, Perú

ppimente@ulima.edu.pe
jcveliz@pucp.edu.pe

Abstract: This article examines how generative artificial intelligence (GenAI) systems reflect pre-existing gender biases and explores their implications in social and technological contexts. While GenAI holds transformative potential in healthcare, employment, and finance, it also poses considerable risks concerning diversity in training data, development teams, and other factors that can reinforce stereotypical representations and discriminatory decisions, highlighting the need for a comprehensive approach to mitigate these issues. The study employs a systematic literature review following PRISMA guidelines, complemented by thematic analysis to identify key patterns. Articles published between 2020 and 2024 were reviewed, focusing on the nature, origins, and implications of gender biases in GenAI. The thematic analysis enabled the identification of emerging trends and proposed solutions, providing a comprehensive view of current limitations and priority areas for future research. The findings reveal that gender biases in GenAI manifest at various levels, ranging from algorithms reinforcing stereotypes to underrepresentation in generated images. The implications include the reinforcement of social inequalities and the erosion of user trust in GenAI systems. However, strategies such as diversifying development teams, using representative datasets, designing equity-aware algorithms, and establishing robust regulations are highlighted as ways to address these challenges. This article contributes to academic and professional fields by offering a detailed analysis of gender biases in GenAI, identifying practices and strategies to build unbiased systems. Furthermore, it emphasizes the importance of raising public awareness and fostering education on gender biases in GenAI to create more critical and informed users.

Keywords: GenAI, Gender Biases, Algorithmic Fairness, Diversity in AI, Technological Stereotypes

1. Introduction

The rise of Generative Artificial Intelligence (GenAI) has intensified concerns about gender bias, a pervasive issue rooted in the data and algorithms that shape these systems. AI models, particularly in natural language processing (NLP), often reflect and amplify societal gender inequalities (Locke & Hodgdon, 2024; Currie et al., 2024). Research shows that even gender-neutral AI applications, such as search engines, frequently produce male-dominated results (Antonopoulou, 2023). These biases stem from training datasets that lack diversity, leading to skewed representations of gender roles (De Silva and Alahakoon 2022; Ferrara, 2024). The impact extends beyond representation, influencing societal perceptions and professional opportunities (Biswas, et al., 2024; Abdelhay et al., 2024). While mitigation strategies like adversarial learning show potential, they fail to eliminate bias entirely (Cirillo and Rementeria, 2022; Leavy et al., 2020). Addressing gender bias requires a multifaceted approach that considers intersectionality (Hall and Ellis 2023), ensuring AI development promotes fairness and inclusivity. This research addresses: What is the current state of research on gender bias in Generative AI, and what challenges persist in mitigating its impact?.

2. Methodology

To ensure a thorough and current analysis, this review examines literature published between 2020 and 2024. Adhering to PRISMA guidelines, a systematic and structured search was conducted. A keyword-based query in the Scopus database identified peer-reviewed articles in English, using terms such as (Generative Artificial Intelligence OR Generative AI AND Gender Bias OR "Algorithmic Bias" OR "Fairness in AI"). The selection criteria prioritized studies within social sciences and technology, excluding those outside these fields or lacking empirical or theoretical contributions. A full-text analysis was performed on the shortlisted articles to verify their relevance, ultimately leading to the inclusion of 42 studies. The complete methodological framework is illustrated in Figure 1.

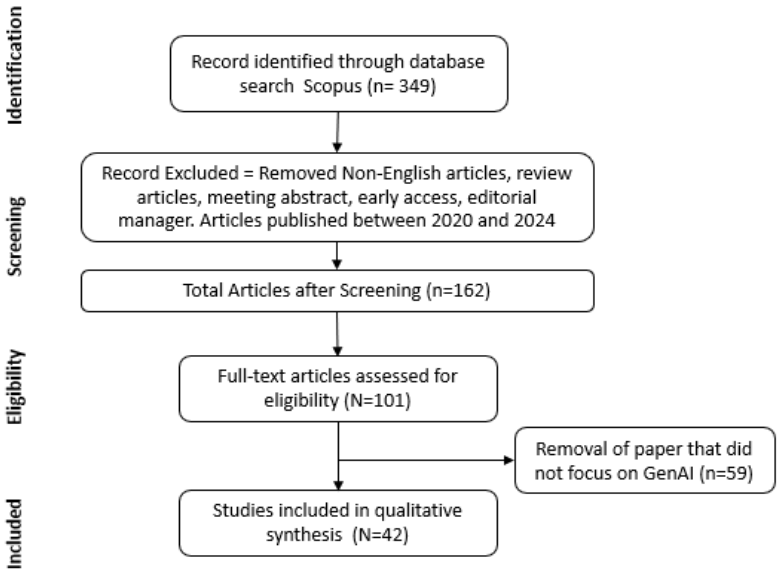

Figure 1: Flow Diagram Prisma

3. Findings

GenAI systems reinforce gender biases across employment, finance, healthcare, and media due to algorithmic design flaws and biased training data. In the job market, AI-powered hiring tools and large language models (LLMs) associate male-coded terms with high-paying roles, systematically disadvantaging female candidates and sustaining economic disparities (Leong and Sung, 2024). Furthermore, AI-generated content in media and digital platforms underrepresents women in male-dominated professions while overrepresenting them in caregiving and administrative roles, reinforcing occupational segregation (Sun et al., 2024; Locke and Hodgdon, 2024). Similar biases exist in finance, where credit-scoring algorithms trained on historically biased datasets restrict women's access to loans and investment opportunities, worsening economic inequality (Ferrara, 2024; Gross, 2023). These biases are further exacerbated by algorithmic opacity, which limits users' ability to contest or even detect discriminatory outcomes, highlighting the need for greater transparency in AI decision-making (Pérez-Ugena Coromina, 2024).

Beyond economic disparities, gender bias in GenAI has significant implications for healthcare. AI-driven diagnostic tools and mental health applications demonstrate lower accuracy for female patients, leading to disparities in treatment recommendations and healthcare accessibility (Park et al., 2022). For example, mobile mental health assessments have shown differential accuracy rates based on gender, disproportionately misdiagnosing female users (Isaksson, 2024). These biases extend to AI-generated medical research content, where male-dominated datasets contribute to underrepresentation in clinical trials, impacting women's health outcomes (Nwafor, 2024). Moreover, GenAI recruitment algorithms, already biased against women, show an even stronger bias against women of color, amplifying the intersectionality of gender discrimination with race and socioeconomic status (Kim et al., 2024). This phenomenon, observed in multiple domains, underscores the necessity of inclusive training datasets that adequately represent diverse demographic groups (Parsheera, 2018).

The social implications of these biases are profound, shaping public perceptions and reinforcing structural inequalities. Gendered AI-generated content perpetuates stereotypes that influence career choices, financial stability, and healthcare access, creating a self-reinforcing cycle of disadvantage (Møgelvang et al., 2024). These biases also impact user trust, as demonstrated by research indicating that individuals—particularly women—express greater skepticism towards AI-generated content due to perceived bias and misinformation (Moon, 2024). Additionally, biased AI systems erode public trust, leading to algorithmic aversion and skepticism toward automated decision-making, particularly among marginalized groups (Zlateva et al., 2024). Addressing these systemic issues requires proactive interventions, including comprehensive bias audits, regulatory oversight, and the implementation of fairness-aware AI models (Newstead et al., 2023). A multi-stakeholder approach involving policymakers, AI developers, and impacted communities is essential to ensuring equitable and transparent AI-driven decision-making (Kim et al., 2024). By integrating diverse perspectives into AI development, it is possible

to mitigate bias and promote fairer technological advancements that do not disproportionately disadvantage specific gender groups.

4. Expected Results and Future Plan

Mitigating gender bias in GenAI requires integrating gender-aware methodologies, diverse datasets, and continuous monitoring to ensure equitable representation and user trust (Zhou et al., 2024; Gross, 2023). While efforts remain inconsistent across linguistic and cultural contexts, successful real-world applications highlight effective strategies. Bias-aware dataset curation has reduced gender disparities by ensuring diverse training data, particularly in AI-generated content where underrepresentation skews outputs (Isaksson, 2024). In healthcare, including marginalized groups in clinical datasets has improved diagnostic accuracy, addressing gender-driven medical bias (Isaksson, 2024). Algorithmic transparency, enforced through systematic audits and bias detection mechanisms, has also proven essential in preventing discriminatory AI outputs before deployment (Nwafor, 2024). Moreover, interdisciplinary collaboration between AI developers, policymakers, and social scientists fosters inclusive governance frameworks that embed fairness principles at every stage of AI development (Isaksson, 2024). Sustainable implementation demands integrating these measures into AI regulations, mandating continuous auditing and proactive intervention to prevent algorithmic discrimination (Nwafor, 2024). Beyond textual and decision-making applications, gender bias also affects AI-generated visual and multimedia content, reinforcing harmful stereotypes in digital media (Zhou et al., 2024). Future research should adopt a cross-cultural and longitudinal approach to assess the persistence and evolution of bias across different regions and over time. Expanding studies geographically will uncover regional disparities, while long-term evaluations will determine the sustained effectiveness of mitigation strategies, ensuring that AI systems evolve to be fair, unbiased, and socially responsible (Muller et al., 2023).

References

Abdelhay, S., Haider, S., Hazaimeh, H. M., El-Bannany, M. y Marie, A. (2024) 'The impact of Generative AI (ChatGPT) on the HR functions related hiring', en *Artificial Intelligence, Digitalization and Regulation: A Legal Framework for Business*, vol.369.

Antonopoulou, C. (2023) 'Algorithmic bias in anthropomorphic artificial intelligence: Critical perspectives through the practice of women media artists and designers', *Technoetic Arts: Journal of Speculative Research*, vol.21, no.2, pp.157-174.

Biswas, S., Jung, J. Y., Unnam, A., Yadav, K., Gupta, S. y Gadiraju, U. (2024) 'Hi. I'm Molly, your virtual interviewer! Exploring the impact of race and gender in AI-powered virtual interview experiences', en *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol.12, pp.12-22.

Cirillo, D. y Rementeria, M. J. (2022) 'Bias and fairness in machine learning and artificial intelligence', en *Sex and Gender Bias in Technology and Artificial Intelligence*, pp.57-75. Academic Press.

Currie, G., Currie, J., Anderson, S. y Hewis, J. (2024) 'Gender bias in generative artificial intelligence text-to-image depiction of medical students', *Health Education Journal*, vol. 83, no. 7, pp. 732-746.

De Silva, D. y Alahakoon, D. (2022) 'An artificial intelligence life cycle: From conception to production', *Patterns*, vol.3, no.6.

Ferrara, E. (2024) 'Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies', *Sci*, vol.6, no.1, p.3.

Gross, N. (2023) 'What ChatGPT tells us about gender: A cautionary tale about performativity and gender biases in AI', *Social Sciences*, vol.12, no.8, p.435.

Hall, P. y Ellis, D. (2023) 'A systematic review of socio-technical gender bias in AI algorithms', *Online Information Review*, vol.47, no.7, pp.1264-1279.

Isaksson, A. (2024) 'Mitigation measures for addressing gender bias in artificial intelligence within healthcare settings: A critical area of sociological inquiry', *AI and Society*, pp.1-10.

Kim, S., Oh, P., and Lee, J. (2024). 'Algorithmic gender bias: investigating perceptions of discrimination in automated decision-making',*Behaviour & Information Technology*, vol.*43,no.*16, pp. 4208-4221.

Leavy, S., Meaney, G., Wade, K. y Greene, D. (2020) 'Mitigating gender bias in machine learning data sets', en *Bias and Social Aspects in Search and Recommendation: First International Workshop, BIAS 2020, Lisbon, Portugal, April 14, Proceedings 1*, pp.12-26. Springer International Publishing.

Leong, K. y Sung, A. (2024) 'Gender stereotypes in artificial intelligence within the accounting profession using large language models', *Humanities and Social Sciences Communications*, vol.11, no.1, pp.1-11.

Locke, L. G. y Hodgdon, G. (2024) 'Gender bias in visual generative artificial intelligence systems and the socialization of AI', *AI and Society*, pp.1-8.

Møgelvang, A., Bjelland, C., Grassini, S. y Ludvigsen, K. (2024) 'Gender differences in the use of generative artificial intelligence chatbots in higher education: Characteristics and consequences', *Education Sciences*, vol.14, no.12, p.1363.

Moon, S. J. (2024) 'Effects of perception of potential risk in generative AI on attitudes and intention to use', *International Journal on Advanced Science, Engineering & Information Technology*, vol.14, p 5.

Muller, B., Alastruey, B., Hansanti, P., Kalbassi, E., Ropers, C., Smith, E. M. y Costa-Jussà, M. R. (2023) 'The Gender-GAP Pipeline: A Gender-Aware Polyglot Pipeline for Gender Characterisation in 55 Languages', *arXiv preprint*, arXiv:2308.16871.

Newstead, T., Eager, B. y Wilson, S. (2023) 'How AI can perpetuate–or help mitigate–gender bias in leadership', *Organizational Dynamics*, vol.52, no.4, p.100998.

Nwafor, I. E. (2024) 'Gender mainstreaming into African artificial intelligence policies: Egypt, Rwanda and Mauritius as case studies', *Law, Technology and Humans*, vol.6, no.2, pp.53-68.

Park, J., Arunachalam, R., Silenzio, V. y Singh, V. K. (2022) 'Fairness in mobile phone–based mental health assessment algorithms: Exploratory study', *JMIR Formative Research*, vol.6, no.6, p.e34366.

Parsheera, S. (2018, November). A gendered perspective on artificial intelligence. In *2018 ITU Kaleidoscope: Machine Learning for a 5G Future (ITU K)* (pp. 1-7). IEEE.

Pérez-Ugena Coromina, M. (2024) 'Sesgo de género (en IA)', *Eunomía. Revista en Cultura de la Legalidad*, vol.26, pp.311-330.

Sun, L., Wei, M., Sun, Y., Suh, Y. J., Shen, L. y Yang, S. (2024) 'Smiling women pitching down: Auditing representational and presentational gender biases in image-generative AI', *Journal of Computer-Mediated Communication*, vol.29, no.1, p. zmad045.

Zhou, H., Inkpen, D. y Kantarci, B. (2024) 'Evaluating and mitigating gender bias in generative large language models', *International Journal of Computers Communications & Control*, vol.19, p.6.

Zlateva, P., Steshina, L., Petukhov, I., and Velev, D. (2024). 'A conceptual framework for solving ethical issues in generative artificial intelligence' In *Electronics, Communications and Networks.*pp. 110-119. IOS Press.