Bias and Discrimination in Artificial Intelligence: A Literature Review

Michael Christopher – S224830467

In high-stakes decision-making processes like recruiting, banking, criminal justice, and healthcare, artificial intelligence (AI) technologies are being utilized more. Concerns regarding bias and discrimination have grown because of the implementation of these systems, especially when AI algorithms unfairly dislike people based on their socioeconomic background, gender, or race. Developing just and moral technology requires an understanding of the ways in which bias enters AI systems. This review of the literature examines the causes, expressions, and suggested solutions of bias and discrimination in AI, utilizing the body of existing research to identify important themes, difficulties, and areas for further investigation.

Artificial intelligence (AI) bias and discrimination are widespread concerns that affect a variety of industries, including recruiting, criminal justice, healthcare, and more. One of the primary concerns is offered by Yavuz (2019), who argues that because the data used to train machine learning systems frequently reflects historical discrimination, the algorithms themselves inherit societal biases. Nelson's (2019) work shares these concerns, noting that skewed data, opaque algorithms, and a lack of development monitoring result in systemic bias, especially in medical AI applications. To provide a formal understanding of how bias appears in different systems, Kundi et al. (2023) expand the scope by conducting a scoping review and classifying AI bias into representational, measurement, and algorithmic forms. Heinrichs (2022) covers the ethical consequences of AI bias in a similar manner, emphasising that discrimination in algorithmic decisions is not only technical but also normative and needs to be considered from a moral and legal perspective.

Tischbirek (2019) adds a regulatory dimension, while Miller (2020) adds a philosophical one. While Tischbirek (2019) argues for the regulation of discriminatory AI systems through anti-discrimination law, placing emphasis on the importance of treating biased algorithms as potential violators of legal norms, Miller (2020) views AI bias as an issue of social inequality, where algorithms mimic existing hierarchies. According to Bagaric et al. (2022), who suggest transparent and equitable AI systems as a solution for criminal justice bias, AI can be morally advantageous provided it is subject to strict design and monitoring guidelines.

Moss (2020) and Miasato and Silva (2019) investigate how biased training datasets and incorrect proxy variables often lead to AI-powered recruiting systems discriminating against people with disabilities and marginalized communities in the workplace. In their detailed review of the AI bias research in human resources management, Kekez et al. (2025) bring up

conceptual ambiguity and a lack of workable mitigation techniques. Additionally, they highlight how some demographics, such as Western, white, and male, are overrepresented in developer teams and datasets, which promotes systematic exclusion. Chen et al. (2023), on the other hand, support human-centered design as a proactive step to incorporate equality and transparency into AI systems from the beginning. Their study offers actual evidence of how interdisciplinary teams and active methods lessen bias throughout the design phase.

Although the research areas differ, they all agree that bias frequently results from socio-technical relationships, which are the intersections of technical design decisions, data practices, and social inequality (Pulivarthy & Whig, 2025). But they differ in the solutions they provide. While Chen et al. (2023) and Kundi et al. (2023) stress structural reforms in AI development pipelines, Heinrichs (2022) and Tischbirek (2019) concentrate on legal and ethical governance. This disparity highlights a larger gap between technical and normative solutions to AI bias.

Although there isn't much experimental support for these claims, some studies, like Chen et al. (2023), present findings from system audits and participatory workshops that show inclusive design techniques produce more equitable results. The over-reliance on theoretical or case-based analysis and the lack of extensive empirical validation, however, are common limitations. Furthermore, although Bagaric et al. (2022) and Moss (2020) offer comprehensive sector-specific solutions, contextual differences in data and decision-making procedures may prevent their methods from generalizing across domains.

The literatures generally follow a path from early diagnostic critiques (Yavuz, 2019; Nelson, 2019) to more complex, multidisciplinary frameworks that consider technical, social, and legal aspects (Pulivarthy & Whig, 2025; Kekez et al., 2025). This development shows that the field is maturing, but there are still many obstacles to overcome before justice, accountability, and transparency can be operationalized across the wide range of AI applications. The body of research highlights that tackling AI bias necessitates structural, legal, and conceptual changes in the way we develop, implement, and govern intelligent systems, more than just algorithmic solutions.

References:

Bagaric, M., Svilar, J., Bull, M., Hunter, D., & Stobbs, N. (2022). The solution to the pervasive bias and discrimination in the criminal justice system: transparent and fair artificial intelligence. *Am. Crim. L. Rev.*, *59*, 95.

Chen, Y., Clayton, E. W., Novak, L. L., Anders, S., & Malin, B. (2023). Human-centered design to address biases in artificial intelligence. *Journal of medical Internet research*, *25*, e43251.

Heinrichs, B. (2022). Discrimination in the age of artificial intelligence. *AI & society*, *37*(1), 143-154.

Kekez, I., Lauwaert, L., & Ređep, N. B. (2025). Is artificial intelligence (AI) research biased and conceptually vague? A systematic review of research on bias and discrimination in the context of using AI in human resource management. *Technology in Society*, 102818.

Kundi, B., El Morr, C., Gorman, R., & Dua, E. (2023). Artificial intelligence and bias: a scoping review. *AI and Society*, 199-215.

Miasato, A., & Silva, F. R. (2019). Artificial intelligence as an instrument of discrimination in workforce recruitment. *Acta Univ Sapientiae: Legal Stud*, *8*(2), 191-212.

Miller, K. (2020). A matter of perspective: Discrimination, bias, and inequality in ai. In *Legal regulations, implications, and issues surrounding digital data* (pp. 182-202). IGI Global.

Moss, H. (2020). Screened out onscreen: Disability discrimination, hiring bias, and artificial intelligence. *DENv. L. REv.*, *98*, 775.

Nelson, G. S. (2019). Bias in artificial intelligence. *North Carolina medical journal*, *80*(4), 220-222.

Pulivarthy, P., & Whig, P. (2025). Bias and fairness addressing discrimination in AI systems. In *Ethical dimensions of AI development* (pp. 103-126). IGI Global.

Tischbirek, A. (2019). Artificial intelligence and discrimination: Discriminating against discriminatory systems. In *Regulating artificial intelligence* (pp. 103-121). Cham: Springer International Publishing.

Yavuz, C. (2019). Machine Bias: artificial intelligence and discrimination.