Combating Bias in Automated Hiring Systems

A Recommendation Report Featuring a Methodological Framework for Fairer Recruitment

Maryam Tariq
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas
Richardson, Texas
maryam.tariq@utdallas.edu

***Abstract*—With the increasing reliance of companies on AI for automated hiring systems, urgent concerns have been raised about the pre-existing biases embedded in these systems. These biases disproportionately affect candidates based on race, gender, and socioeconomic status. This paper delves into the causes of algorithmic bias in these models and its impact on hiring outcomes. It also presents a methodological framework to mitigate bias. This paper draws on primary and secondary sources, including studies and industry reports, outlines both ethical and technical facets of these issues, and recommends a three-pronged strategy.**

Keywords—bias, artificial intelligence, hiring systems, fairness, recruitment, machine learning, automated hiring, discrimination, audits, processes, training data, transparency

I. INTRODUCTION

This paper aims to present a transparent, detailed methodology to combat bias in automated hiring systems. This practical methodology provides a strong recommendation for companies and organizations seeking to avoid discrimination and bias in the hiring process. Automated processes and tracking systems for hiring and recruitment have become an increasingly popular alternative to traditional hiring methods to analyze resumes, rank candidates, and assess video interviews. However, concerns have been raised regarding its amplification and conformity to preexisting biases in hiring decisions. This paper examines the types of bias found in automated hiring systems, analyzes the factors contributing to these biases, and recommends a methodological solution to mitigate and address these biases.

II. TYPES OF BIAS IN AUTOMATED HIRING SYSTEMS

A. Race and gender bias

Automated hiring systems are susceptible to various forms of bias that can disproportionately affect candidates. One of the most prevalent is race and gender bias, often based on historical patterns of discrimination in training data. Gender and race are intersecting systems of power rather than fixed variables. Removing race and gender to promote universal subjectivity and sameness does not allow for a fair review of groups and inclusivity based on a standard candidate [1].

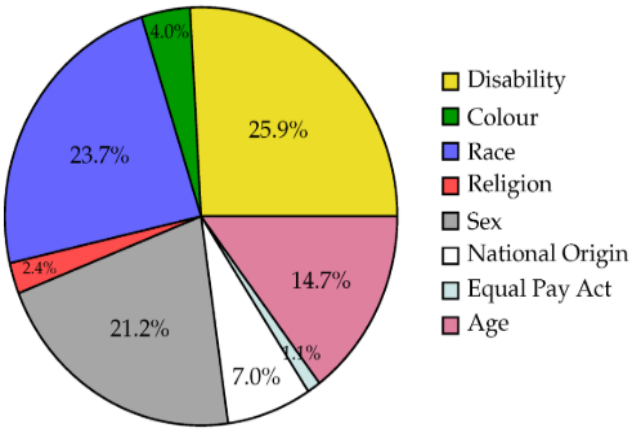

Fig. 1. Causes of Discrimination in the Hiring Process. Source: [3].

As illustrated in Figure 1, race and sex (gender) account for a substantial portion of discrimination in the hiring process—23.7% and 21.2% respectively. This data underscores the impact of racial and gender factors in hiring outcomes, and automated hiring systems that are modeled off biased data risk replication historical patterns of discrimination.

B. Institutional bias

Another form of bias is institutional bias, based on systematic practices and norms that can disadvantage certain groups in algorithmic hiring. For example, horizontal segregation reflects disparities in employment rates across industry sectors associated with sensitive attributes. Likewise, vertical segregation displays disparities in career advancement to leadership positions. When translated into data, it leads to models reinforcing wage gaps and a lack of diversity in higher positions. In one case, social network topology influences the likelihood of being screened by a recruiter due to a successful referral. However, women are less likely to be recommended by a system than men [2].

C. Algorithmic bias

The last form of bias mentioned in this paper is the structural form of algorithmic bias. The measurement bias occurs when training data for algorithms cannot accurately represent the intended construct it seeks to measure. Another form is

representation bias when non-representative samples of the population are used for data, which leads to under-representation and overrepresentation of sample groups. The omitted variable bias is when algorithms have one or more important variables missing in a model, affecting system prediction. Further, the linking bias occurs when user connections misrepresent a candidate's behavior. Lastly, the aggregation bias comprises false conclusions made regarding an individual based on the analysis of the entire population [3].

Together, these multifaceted biases in automated hiring systems illustrate how data and algorithms predict unequal hiring outcomes, even when automated models appear objected.

III. FACTORS CONTRIBUTING TO BIAS

A. Data Related Biases

Data is the foundation of AI hiring systems, and biases directly affect the fairness of algorithms. The first source of bias stems from representation issues or underrepresentation of minority groups and overrepresentation of dominant groups. Suppose a dataset lacks sufficient information from specific demographics (e.g., women, racial minorities, people with disabilities). In that case, the algorithm may be unable to accurately assess these groups, which leads to such candidates being overlooked or misclassified. Likewise, suppose data is skewed towards a specific demographic. In that case, the algorithm may translate those characteristics particular to the group as indicators of success, which disadvantages candidates who do not possess those characteristics. The second source of bias is historical disparities in hiring practices and systemic inequality. Historical hiring data may reflect past discrimination. When trained on these data sets, algorithms amplify and replicate those biases. Further, broader social inequalities, like access to education and opportunity, are reflected in hiring, which can lead to models preferring candidates from privileged backgrounds [4].

B. Feature Selection and Weight Assignment in Algorithms

Feature selection identifies the variables a model uses to evaluate candidates, while weight assignment determines the importance of each variable. Some features (such as zip code or university) can act as proxy variables for protected characteristics (race, gender, socioeconomic status). An algorithm may conclude that alums from elite universities are stronger job performers, which disadvantages qualified candidates from less prestigious schools. Overweighted correlated features and omission of contextual variables can impose patterns of exclusion by penalizing candidates due to individual, context-sensitive circumstances, such as career gaps and non-traditional paths [4].

C. Lack of Transparency

The complex, not easily understandable nature of algorithms means their decision-making processes lack transparency for users and affected individuals. This lack of transparency underscores the need for clear decision criteria to ensure fair outcomes and identify and correct bias. The challenge of holding organizations accountable for biased outcomes or explanations further emphasizes the importance of transparency in AI decision-making [4].

D. Unintended Reinforcement of Human Biases

AI systems can inadvertently reinforce human bias through biased input during model training when defining an ideal candidate can encode subjective preference and prejudice. Feedback loops and automation bias mean that over-reliance on self-operating cycles of bias means the model will continuously favor a particular set of variables and reduce the role of human intervention in ensuring fair evaluation [4].

IV. CASE STUDIES

A. Amazon's Hiring Algorithm

In 2014, Amazon implemented a tool to review job applicant resumes using natural language processing (NLP) and machine learning (ML). This software would then use AI algorithms to learn key qualities from successful candidate resumes and look for similarities in resumes submitted for screening. This tool would then rate the candidate out of 5 stars on their resemblance to prior successful candidates. However, since the technology industry trend is primarily male, biased data was used to train the AI system. As a result, in 2018, it was reported that algorithms in this tool created an association that downgraded resumes that contained "women" in them, providing a case of real-life gender bias in an AI recruiting tool [5].

B. HireVue's Video Interview System

HireVue is an automated hiring system (AHS) that performs automated video interviews to profile candidates and promises to eliminate human bias in assessing candidates based on two strategies using vocal and facial assessments [6]. Firstly, by removing the indicators or features that have an adverse impact on a protected group based on previous knowledge, and secondly, by modifying the learning algorithm for fairness. However, this system has limitations. The system's reliance on historical data can reflect bias as such algorithmic specification to the best fit can become a vehicle for bias in logic prediction. HireVue also uses the 4/5ths rule, a guideline to assess the negative impact on a protected group in hiring, does not account for intersectional discrimination, such as overlapping forms of disadvantage or bias due to multiple identities (black women vs. white men). Lastly, the abstract nature of its validation mechanism and little transparency makes it difficult for audits to assess fairness accurately [7].

V. PROPOSED METHODOLOGICAL RECOMMENDATION

A. Improving Training Data Diversity and Representativeness

One of the best ways to reduce algorithmic bias is to ensure the training data reflects the diversity of a potential candidate pool. To improve data quality, we can enhance representation by including diverse demographic groups in datasets and encapsulating different applicant experiences based on geographic regions, industries, and career paths. We can correct historical bias by removing patterns of discrimination from data by excluding biased outcomes, adjusting for known disparities, and re-sampling to balance the dataset. Synthetic data generation and inclusion data collection ensure that data captures diverse outcomes for more equitable models and inclusive criteria [4].

B. *Developing Fairness Aware Algorithms*

Algorithms with fairness-aware algorithms can correct bias through specific techniques and optimizations. Pre-processing techniques include reweighting data to equalize representation and transforming data to remove sensitive attributes and outcome correlations. In-processing techniques include modifying objective functions and using adversarial models to detect and minimize bias during model learning. Post-processing techniques include adjusting predictions to assure equal decision rates across protected groups and implementing rules of fairness. Multi-objective optimization balances accuracy and fairness for maximized predictive performance [4].

C. *Implementing Explainable AI*

Implementing explainable AI would mean using interpretable models such as decision trees and logistic regression instead of complex black-box models. Feature attribution could identify what most influences hiring outcomes, and transparent reporting provides candidates and hiring managers with clear explanations for algorithmic decisions [4].

D. *Regular Auditing and Monitoring*

Regularly auditing and monitoring systems through bias audits and third-party audits can validate fairness claims and verify compliance with legal and ethical standards. Performance monitoring of models can detect and address emerging bias, and model drift detection can monitor shifts that can potentially reintroduce bias [4].

E. *Similarities in Gender Treatment in the Workplace*

Human review can check and balance automated systems. Human recruiters can make final hiring decisions, and ambiguous and sensitive decisions can be flagged for human review. Training for bias awareness, establishing oversight, and encouraging candidate feedback on the process can reduce and improve transparency. Such transparency led participants to perceive algorithms as free from personal bias that humans could raise in decision-making processes [4].

VI. COUNTERARGUMENT

Proponents claim that when transparency is provided, algorithms can be perceived as more trustworthy and fairer than human recruiters. In a 2023 study where participants were informed on algorithm training and decision-making, such transparency led participants to perceive algorithms as free from personal bias humans could raise in decision-making processes [8].

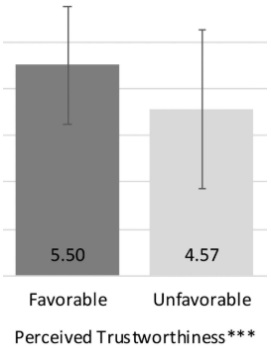

Fig. 2. Mean differences in perceived trustworthiness between favorable and unfavorable outcome conditions. ***p < 0.001. Adapted from [8].

Research shows that algorithmic appreciation can occur even when outcomes are unfavorable. As shown in Figure 2, participants rated algorithms as more trustworthy than humans when the outcome was negative (mean = 4.57, p < 0.001), which indicates there is a preference to algorithmic decision - makers.

However, while transparency can improve perceptions, it alone does not guarantee fairness. Explanation of model logic can create the illusion of objectivity, but as explored in case studies mentioned above, the data and logic underneath can remain biased or unclear. In a 2024 study, applicants' perceptions of procedural and distributive fairness are lower when using AI resume screening than traditional human resource hiring methods [9].

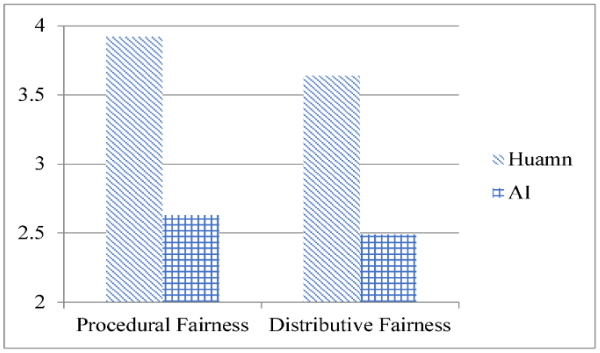

Fig 3. Mean difference between resume screeners on applicants' perceptions of fairness. Source: [9].

While some evidence supports algorithmic appreciation, more recent studies, as shown in Figure 3, reveal that applicants perceived human resume screeners as significantly fairer than AI in both procedural and distributive fairness. This challenges the notion that transparency alone can build trust and ensure fairness in algorithms.

VII. CONCLUSION

Automated hiring systems are designed for efficiency and the objective of recruiting the best candidates. Despite this goal, there are significant risks of bias through flawed data, errors in model logic, and lack of transparency. This bias disproportionately disadvantages qualified candidates based on their background or without considering contextual circumstances.

This paper recommends a three-pronged strategy to mitigate the disadvantages found in automated hiring systems.

1. Improving the diversity and representativeness of training data and applying fairness-aware strategies (reweight, adversarial debiasing, and fairness constraints) to reduce bias throughout the model processing lifecycle.

2. Using interpretable models and post-hoc explanation tools to ensure hiring outcomes are transparent and justifiable to candidates and recruiters.

3. Combining regular auditing and monitoring with human oversight to ensure accountability and ethical compliance.

To move forward, companies and organizations must adopt this methodological framework to take proactive, structural action as automated systems become increasingly embedded in recruitment. Responsible initiative can achieve change at every stage in the development lifecycle. Algorithmic fairness should be treated as a core component of corporate responsibility, not just a marker of legal and ethical compliance, as transparency and trust are industry standards between candidates and recruiters. Future research can explore the refinement of this framework by testing it across diverse hiring platforms, industries, and roles to ensure adaptability and effectiveness. With this operational shift, we can move towards a more equitable landscape in the domain of automated recruitment hiring systems.

REFERENCES

[1] E. Drage and K. Mackereth, "Does AI debias recruitment? Race, gender, and AI's 'eradication of difference'," *Philosophy & Technology*, vol. 35, no. 89, 2022. [Online]. Available: https://doi.org/10.1007/s13347-022-00543-1. [Accessed: May 5, 2025].

[2] A. Fabris, N. Baranowska, M. J. Dennis, D. Graus, P. Hacker, J. Saldivar, F. Zuiderveen Borgesius, and A. J. Biega, "Fairness and bias in algorithmic hiring: A multidisciplinary survey," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 1, Art. no. 16, pp. 1–54, Jan. 2025. [Online]. Available: https://doi.org/10.1145/3696457. [Accessed: May 5, 2025].

[3] E. Albaroudi, T. Mansouri, and A. Alameer, "A comprehensive review of AI techniques for addressing algorithmic bias in job hiring," *AI*, vol. 5, no. 1, pp. 383–404, 2024. [Online]. Available: https://doi.org/10.3390/ai5010019. [Accessed: May 5, 2025].

[4] M. Busari, *Algorithmic bias in hiring automation: Ensuring fairness and diversity in AI-driven recruitment*, 2025. [Online]. Available: https://www.researchgate.net/profile/Muhammed-Busari/publication/390141550_Algorithmic_Bias_in_Hiring_Automation_Ensuring_Fairness_and_Diversity_in_AI-Driven_Recruitment/links/67e2065c72f7f37c3e8ad8d5/Algorithmic-Bias-in-Hiring-Automation-Ensuring-Fairness-and-Diversity-in-AI-Driven-Recruitment.pdf. [Accessed: May 5, 2025].

[5] A. Kodiyan and A. Alfons, "An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI-based hiring tool," *ResearchGate Preprint*, vol. 12, pp. 1-9, 2019. [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/63223371/Essay_On_Ethics___AI_Hiring20200506-124352-1cpydva-libre.pdf?1588823614=&response-content-disposition=inline%3B+filename%3DAn_overview_of_ethical_issues_in_using_A.pdf&Expires=1746317811&Signature=EQjeM~uBYg47FtYFd4zpCV3NFMhjuLQT8iTvpdWC4zEUbkOyRhsWixH291745UXOODOPat~AWI014DBe8c~GYNlW0sxjB4Mng5TKA6K5HggFXFGNsARe4~ROdkyxwCRMGKCVIZHphizGUlOSMvs7DRFUnWLpgPxV3ddMqHajrd1p651lV2CJ8RdDliZosVDzwhFaOu5aZoOs6YuRkg-iDfQSRmxawTjuHCRevPR0IbN7-mvF3cNHrAu7r65pHMCzW0GHv~tY3xr-jwqtWA6jm5qgWDckIo6Jtyxs4sOWbRIyhloGJ-fvbpkzMVVcyCLs-DMiCPdqvuvod~gIUXiOBw__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA. [Accessed: May 5, 2025].

[6] I. Ajunwa, "Automated video interviewing as the new phrenology," *Berkeley Technology Law Journal*, vol. 36, no. 3, pp. 1173–226, 2021. [Online]. Available: https://www.jstor.org/stable/27210405. [Accessed: May 5, 2025].

[7] J. Sánchez-Monedero, L. Dencik, and L. Edwards, "What does it mean to 'solve' the problem of discrimination in hiring?," *SSRN*, 11 pp., Oct. 2, 2019. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3463141. [Accessed: May 5, 2025].

[8] H. Choung, J. S. Seberger, and P. David, "When AI is perceived to be fairer than a human: Understanding perceptions of algorithmic decisions in a job application context," *International Journal of Human–Computer Interaction*, vol. 40, no. 22, pp. 7451–7468, 2023. [Online]. Available: https://doi.org/10.1080/10447318.2023.2266244. [Accessed: May 5, 2025].

[9] F. Cai, J. Zhang, and L. Zhang, "The impact of artificial intelligence replacing humans in making human resource management decisions on fairness: A case of resume screening," *Sustainability*, vol. 16, no. 9, p. 3840, 2024. [Online]. Available: https://doi.org/10.3390/su16093840. [Accessed: May 5, 2025].