

Artificial Intelligence for Fair Hiring: A Critical Review of Diversity and Inclusion Outcomes

Sadique Nayeem*

Abstract

Artificial Intelligence has become one of the most innovative technologies to sweep across industries over recent years and recruitment is no exception. In light of recruitment being powered by Artificial Intelligence technologies, including machine learning algorithms and Natural Language Processing, there is great potential for impacting the future of diversity and inclusion within our workplaces. From automated screening of candidates to sourcing new talent, AI presents many opportunities and challenges when it comes to increasing diversity among your talent pool. Opportunities include limiting human error and bias when assessing candidates by standardizing interviews and hiding candidate screening. Challenges including unconscious bias built into your algorithm or lack of transparency (leading to unconscious bias) can harm your diversity efforts. With technology rapidly changing every day and AI continuing to heat up in the recruitment world, it's important to understand where these opportunities and challenges lie to help move diversity forward in your organization. This paper will highlight examples of both AI-powered recruitment harming (Amazon's AI rejected female candidates) and increasing (Unilever's AI reached 16% more diverse pool of candidates) diversity among candidates to show where we stand with AI today. We will then discuss some opportunities for future research to continue moving AI forward with diversity in mind, such as creating auditing measures, inclusion conscious algorithms, and even regulations.

Keywords: Artificial intelligence (AI), machine learning, recruitment, AI-driven systems, re-skilling

INTRODUCTION

The integration of Artificial Intelligence (AI) into recruitment processes has accelerated rapidly, opening doors to a world of possibilities when it comes to streamlining recruitment, bringing objectivity to the table, and creating a level playing field for all candidates. In particular, recruiters are looking to AI to help create a more diverse and inclusive workforce. For decades, hiring processes have been biased toward certain groups of people. Whether intentional or unconscious, human bias in recruitment has created and maintained gender-, race-, age-, and class-based inequalities. Artificial intelligence has the potential to level the playing field for all candidates through the use of machine learning and natural

*Author for Correspondence

Sadique Nayeem

E-mail: sadiquenayeem@gmail.com

Assistant Professor, Department of Computer Science and Engineering, Sitamarhi Institute of Technology, Dumra Sitamarhi, Bihar, India

Received Date: December 26, 2025

Accepted Date: May 13, 2026

Published Date: May 15, 2026

Citation: Sadique Nayeem. Artificial Intelligence for Fair Hiring: A Critical Review of Diversity and Inclusion Outcomes. Journal of Artificial Intelligence Research & Advances. 2026; 13(2): 9–15p.

language processing (NLP) to standardize candidate assessments, anonymize applicant information, and help recruiters identify high-performing candidates from underrepresented pools of talent. Despite this promise, there is an ongoing battle surrounding the ethics and efficacy of using AI to create a fair and just recruiting process. Bias, transparency, and ethics are at the forefront of many people's minds when thinking about AI and diversity. Many are asking, "Does AI help create a more inclusive workplace or further entrench?" Businesses across the globe are making diversity and inclusion (D&I) a top business initiative. Many studies have shown

how diversity can benefit a business. Companies with diverse teams have been shown to be more innovative, have happier employees, and generate higher revenues. However, despite these efforts, underrepresented groups are still being left out. AI has the potential to help alleviate human biases in recruitment by objectively evaluating resumes, standardizing interview questions, and assessing skills rather than educational backgrounds. Here are a few ways AI can help create a more diverse and inclusive workforce. By hiding candidates' names, gender, and age from recruiters, AI has the ability to help level the playing field for all job applicants. Similarly, AI can help level the educational background playing field by placing value on skills rather than the prestige of educational institutions. Another way AI can promote a diverse and inclusive workplace is by helping companies reword biased job descriptions. While AI has the power to create a more diverse workplace, it is important to note that just as AI can remove human bias, it can also leak it. The AI recruitment tool was scrapped after it showed bias against female candidates.

AI algorithms must be fair and explicable. Many algorithmic systems used in hiring are not black boxes, meaning that recruiters and applicants cannot see how they make decisions. This lack of explainability can cause people to lose faith in the system, especially among historically marginalized groups that are rightfully skeptical of AI making decisions about them. Research also indicates that recruiters are likely to distrust algorithms, even when they know they are more effective than human judgement at predicting performance, because they view human decision-making as fairer. There is also a risk that machine learning algorithms may learn the biases present in past hiring decisions. For instance, if a company had historically hired primarily from specific schools or backgrounds, a machine learning model could pick up this pattern and continue it without outside intervention. Some techniques to address this issue involve bias-mitigation techniques, such as adversarial de-biasing, incorporating fairness constraints, or directly diversifying training data. While these processes can limit bias, they can also reduce the overall accuracy of the selection system if not implemented carefully. HR and data science teams must work together with diversity and inclusion leaders to understand how to build AI hiring tools that work for the organization.

Humans are still responsible for working with AI tools. They can screen resumes at a fraction of the time it would take a human, but there is still a human element in hiring. AI will never be able to fully take over this process; there will always be a human element to determine cultural fit and ensure everything is okay on the ethics and fairness level. Research shows that AI will not be as effective at reducing bias and increasing diversity if recruiters only use it for resume screening. AI allows recruiters to create shortlists quicker, but if the recruiter then picks who they want to interview based on their own biases, then they are back to square one. AI literacy training for HR professionals can help them better understand and work with AI tools.

Acknowledging these considerations, this literature review covers how AI can be used to benefit diversity and inclusion when recruiting. We cover current research on algorithmic fairness and debiasing methods, human-AI collaboration, and challenges and limitations, as well as provide ethical and practical considerations. Additionally, we identify areas for future research. These areas include developing standard bias audit procedures, regulations, and governance, and applying fairness approaches to create culturally aware AI. By examining the potential benefits and risks of using AI in hiring, we can work towards creating a responsible and inclusive future for hiring. At the end of the day, if we are to successfully implement AI in recruitment, we must learn how to balance the use of technology with human oversight to establish hiring practices that are fair for all.

Review of Different Research Papers

Shneiderman (2020) [1] proposed a two-dimensional matrix for Human-Centered Artificial Intelligence (HCAI) based on the independent dimensions of human control and computer automation. Previous automation frameworks, such as those proposed by Sheridan and Verplank, suggest that there is a trade-off between automation and human control. However, advances in human-centered AI technology allow increased reliability, safety, and trustworthiness (RST) by balancing human control

and automation to assist with effective human performance, focusing on self-efficacy, creativity, and responsibility. Shneiderman presents four quadrants of high human and low automation (human mastery), low human and high automation (computer autonomy), low human/low automation, and the optimal and desirable upper-right quadrant [1]. There is a need for human-centered AI that balances human control and automation, as both too little automation (e.g., manual piloting of planes) and too much automation (the recent Boeing 737 MAX crashes) have led to fatal mistakes. Additionally, excessive human control can lead to human error. Shneiderman used examples of elevators, digital cameras, and patient-controlled analgesia (PCA) to highlight useful technologies that have achieved RST [1]. To achieve RST, Shneiderman describes various technical practices for reliability, organizational approaches to openness for safety, and independent oversight to gain trustworthiness [1]. Technologies such as self-driving cars are heading down an unrealistic path to 100% autonomy and should instead focus on incremental advancements. The Prometheus Principles should be used as guidelines for interfaces to inform and respond to users' intent, allow for feedback, and prevent errors. Shneiderman (2020) concluded with a plea for researchers to stop developing autonomous AI and start researching and developing HCAI that focuses on both human control and automation [1].

In Tessa Haesevoets et al. (2021) [2] article, “Should Machines Advise or Decide? Mapping Managers’ Preferences for Human versus Machine Involvement in Decision Making, the authors conducted five studies interviewing a total of 1,025 managers. They concluded that “overall, people have a preference for human-machine collaboration with humans in the majority position”. Their findings showed that acceptance rates increased linearly only until machines decided more than humans (70% vs. 30%) and plateaued thereafter. Additionally, they outlined four clusters of managers: 5% would prefer machines to be in the majority, 15% would prefer equal involvement, 50% would prefer human-machine collaboration with humans in the lead, and 30% would only accept humans in the decision-making process. While each study further investigated the topic of AI involvement and what factors influence people’s decision to trust the algorithm, the concluding takeaways are that people have a great variance in their comfort level with AI and that there is a user need for human involvement. As such, AI should not be seen as a replacement but as a “machine-in-the-loop” and people’s preferences should be investigated when implementing AI into the workplace. One major way to do so could be highlighted by one of the studies that showed that people are more prone to trust the algorithm if it is explainable. Therefore, organizations should put effort into training those who are less prone to accepting automated decisions.

Lacroux et al. (2022) [3] investigated recruiters’ trust and responses towards algorithm-based recommendation systems (ADSS) used for screening resumes and their reactions to algorithmic recommendations versus recommendations provided by human experts. This study was based on a sample of 694 professionals who performed a task-based randomized experiment with a 2×2 factorial design. Participants were presented with either consistent or inconsistent recommendations, and the source was either an algorithm or a human. The results indicated that participants rated recommendations made by humans as more trustworthy than algorithm-based recommendations (supporting H1a). When receiving inconsistent recommendations from the algorithm, recruiters changed their decisions and chose unsuitable resumes more often than when they received inconsistent recommendations from human experts (not supporting H2b). The researchers reported that 55.4% of participants chose unsuitable resumes when they received inconsistent advice from the ADSS, while only 40% of participants did so when they received inconsistent recommendations from humans. Additionally, participants’ personalities had a small effect on their trust in ADSS and humans; extraversion was positively correlated with trust in ADSS, and neuroticism had a negative correlation with trust in humans and ADSS. The authors concluded that recruiters’ responses were not aligned with their claims regarding their trust in humans versus algorithms, as despite claiming to trust humans more, they relied on algorithm recommendations. Therefore, recruiters appear to be affected by automation bias. The findings of this study can improve HR policies regarding collaboration between AI and HR professionals. This study can serve as a starting point for additional research regarding building trust in

algorithms, as recruiters did not trust ADSS despite providing better recommendations and the black box model of ADSS.

Edmond et al. (2023) [4] tested how AI affects applicant behavior and the consequences of AI or de-biasing on quality and gender diversity. Using a field experiment simulating the labor market, participants applied for competitive jobs that were randomly assigned to be reviewed by either a human evaluator or an AI evaluator (both standardized and de-biased), using an online platform and a representative sample of the US population (N=1,002). They found that "the use of conventional AI interviewers does not increase diversity or quality relative to human interviewers." However, when evaluators were systematically debiased (for both AI and humans), the participants were much more diverse but equally qualified. For example, compared to human evaluators without de-biasing training, significantly more women applied for high-paying competitive jobs when evaluators were de-biased (66.1% of men and 62.6% of women compared to 75.7% of men and 40.6% of women). The study additionally finds that women applicants strongly prefer de-biased evaluators because they believe they will be treated more fairly and that performance will matter most.

Will et al. (2023) [5] published an article that discusses whether AI recruitment measures are comparable to human recruiters. The authors present a framework titled HIRE, an acronym for Human, (Artificial) Intelligence, Recruitment, and Evaluation, that examines four themes in recruitment: efficiency, performance, diversity, and perceptions. The authors systematically searched and reviewed 22 empirical articles to evaluate how AI measures up against humans on each of the aforementioned themes. They found that AI was either on par or superior to human recruiters when analyzing efficiency and performance. Regarding efficiency, algorithms have been shown to speed up the hiring process. Regarding performance, algorithms were found to predict job performance better than recruiters. One study showed that algorithms increased hiring rates by 20% for technical positions, and AI recruiters selected candidates with higher predicted job performance than human recruiters. Regarding diversity, AI was less biased than humans. The use of algorithms led to a 24% increase in minority hiring in some studies. The effect of AI on diversity differed depending on the type of algorithm used. Under the perception theme, the authors discovered that AI was perceived more negatively by both candidates and recruiters. Candidates were less likely to apply if they knew that AI was being used throughout the recruitment process. They noted that AI was scored significantly lower on qualities such as fairness, trustworthiness, and personability against recruiters. An interesting phenomenon called algorithm aversion occurs when people tend to disregard the use of algorithms despite their higher performance compared to humans. The authors call for more research on the intersectional effects of AI and its implementation in real-world settings. Future research should use the HIRE framework to organize and report results. AI has better efficiency and performance than human recruiters but is perceived poorly by candidates and recruiters.

Wilkens et al. (2023) [6] specified human-centricity based on eight criteria. They allocated eight criteria into three themes: challenges technology should consider (trustworthiness and explainability), workplace challenges (job loss/job deprivation and health/human agency), and organizational challenges (system vulnerabilities, user-domain expertise, and accountability). They gathered 101 articles from a systematic literature review. They used their gathered criteria to create seven unique configurations of actor-structure engagements that highlighted one prominent criterion while including other stakeholders. For example, one mentioned that stakeholders such as AI developers prioritize Data and Techno-structure, while HR Personnel/Staff focus on employee well-being (health). This study showed that all seven configurations of workplace engagement required trustworthiness to be complete. Each configuration showed that academia included many norms that industry standards lack, such as human agency and accountability of the organization. This study can be improved by completing the QCA to empirically test its configurations.

Bursell et al. (2024) [7] describe efforts to scale AI hiring at a large Swedish retailer, pseudonymously named "FoodMarket." The authors develop the notion of meta-algorithmic judgment to help understand

recruiters' interactions with algorithms at scale. Field experiments comparing traditional versus AI hiring practices were directly tested against employment records ($n = 12,336$ hires) and job application data ($n = 187,880$ observations). Unexpectedly, AI hiring processes resulted in less diversity of hires in age, gender, and perceived ethnicity of candidate names. Meta-judgment was biased towards younger workers and moved away from diversity initiatives. On average, algorithm-driven hiring raised the probability of hiring European-named candidates by 7 percentage points relative to the average and raised the probability of hiring women by 6 percentage points relative to traditional processes. Interestingly, most hiring biases did not originate from the algorithm itself. Even though women performed better on algorithmic assessments administered prior to interviews (women: 74.6 vs. men: 70.3 on the algorithmic test), recruiters overwhelmingly invited people with European names to audition at later stages of hiring (84% of all hires). Recruiters deviated from algorithmically generated rankings 70% of the time when interviewing candidates, essentially reintroducing the bias that AI systems had eliminated. These results demonstrate how large-scale AI hiring can lead to unintended interaction effects between humans and algorithms because of recruiters' misunderstanding of or non-compliance with automated systems. Measures to improve cooperation between algorithmic decision-making systems and organizations may improve outcomes, such as organizational monitoring and training of staff involved in hiring.

Bano et al. (2024) [8] discussed a study on how values around diversity and inclusion (D&I) could be embedded across AI recruitment systems. This paper discusses the development of a co-design workshop facilitated in collaboration with SEEK, an online employment platform. Automating recruitment techniques using AI will allow swifter and more effective hiring processes. However, there is a danger of replicating bias through algorithmic decision-making systems that harm marginalized and minority groups. One high-profile example is Amazon's discontinuation of its AI hiring tool after discovering that it was biased against women. This study concerned two AI use cases within SEEK: "Strong Applicant Badge" and "Ads Summarization." Employees created user stories to consider the sources of bias and other barriers to D&I for each use case. Several examples participants discussed included maintaining the proportional representation of minorities for badges and ethnocentrism/language barriers around job advertisements. After the workshop, the participants were interviewed about their experiences and changes in perception. While participants noted an increased awareness of D&I risks, they still felt difficulty in acting on these values. Participants cited limitations in demographic data and the lack of priority from leadership as barriers to promoting D&I. The important takeaways from this study include the need for contextualized guidelines around D&I standards based on the use case, continued buy-in from stakeholders involved in AI implementation, and top-down support to make these values routine.

Pant et al. (2024) [9] discussed how Artificial Intelligence (AI) impacts and changes the way we think about diversity and inclusion (D&I) from an HR perspective. The author explains how machine learning algorithms and natural language processing tools help mitigate human bias that often occurs during hiring processes, performance reviews, and training and development. Using AI-enabled platforms, companies such as Unilever have been able to reduce bias in talent acquisition by 16% through training AI with unbiased datasets and human intervention. The benefits of implementing AI in the workplace include faster recruitment through the automation of repetitive tasks, access to a wider talent pool, and data-driven decision-making. The article briefly touches on some challenges AI creates for diversity and inclusion, such as lack of transparency, algorithmic biases, and data privacy. An example of algorithmic bias is provided through Amazon's scrapped AI recruiting tool that favored men over women for job opportunities. The tool was fed past recruitment data, which was mostly male, causing the AI to disproportionately favor résumés that matched the demographic of past candidates. The ethical considerations regarding AI include anti-discrimination laws, responsible AI usage, and training AI with diverse datasets. Support for these claims is provided through company case studies, such as IBM's ethical AI principles and HireVue's assessment partnerships. The author concludes that AI has the potential to create great change for diversity and inclusion but is only as good as the people

programming it. The future of AI in D&I could include the detection of biases and stereotypes in real time and more specific regulations that AI companies need to follow.

Zhang et al. (2024) [10] conducted a study on the use of Artificial Intelligence (AI) in recruitment and selection among HR professionals in Company A, otherwise known as CATSearch HR Consultancy Inc., and Company B Ventus Search Asia, located in the Philippines. The study utilized the Unified Theory of Acceptance and Use of Technology (UTAUT) as the foundation of their research, which aims to determine the factors that encourage individuals to adopt AI technology for recruitment. A quantitative descriptive method was used, and a structured survey was distributed to 119 HR professionals. The results showed that education influenced the participants attitude towards adopting AI in recruitment, where 86.6% of the participants had a bachelor's degree. Performance expectancy (mean =2.76) and Effort Expectancy (mean =2.72) had positive scores, and social influence (mean =2.72) influenced Company B more than Company A. Regression findings suggested that performance expectancy and effort expectancy were the predictors of behavioral intention, while age, gender, and experience did not significantly predict the model. Several implications were suggested where AI could help recruitment bring better efficiency at work, help screen better qualified applicants, reduce personal bias, save time, and improve customer experience. However, there are still limitations to AI, such as ethical issues, transparency of the process, technical difficulties, invalid information, and biased information. Recommendations for both companies include continuous assessments and training for HR professionals regarding AI technology, ethical practices when programming AI, and strengthening communication between the two departments.

Wilkens et al. (2025) [11] asked under which conditions AI-based decision support tools could help decision-makers hire more diversely and augment their capabilities. To answer this question, the authors conducted an experiment in which 139 decision-makers made 278 hiring decisions in the medical domain. Their main independent variables were whether AI decision tools provided diversity-related explanations, whether hiring was portrayed as quality-sensitive (high-stakes job) or not, and whether the hiring organization provided explicit diversity guidelines. AI decision aids without additional features did not lead to more diverse decisions. Diversity-related explanations increased the odds of choosing a female candidate by 154%. Hiring for quality-sensitive positions and having explicit diversity guidelines increased the odds by 437% and 308%, respectively. AI-literacy and HR-related experience of decision-makers had no significant effect. Based on qualitative coding of decision-makers' reasoning, the decision path towards augmented decision-making (reasoning more reflectively) was more likely if multiple decision-enhancing conditions were given. This finding builds on sociomaterial theorizing and demonstrates how technology, contextual features, and users interact. From a practical standpoint, human resource managers and AI developers can implement certain design features, such as explainability functions and diversity guidelines, to improve diversity in hiring. This study makes several contributions to the discussion of AI's role in HRM by providing contextual factors that can help companies achieve their diversity goals. Limitations include only moderate amounts of explained variance and a lack of qualitative and long-term investigations in the literature.

CONCLUSION

Overall, the primary themes that resonated across all reviewed papers included the necessity for AI-enabled machines to function in human-centered systems to make optimal decisions and how AI can help achieve diversity in recruitment. Simultaneously, significant limitations and ethical considerations have been discussed in the literature. AI that allows for human-in-command yielded more reliable and safe systems, while AI-enabled managers were preferred over AI-autonomous managers. There is a great need for explainability and transparency to foster trust among users. The use of AI for recruitment allows for better efficiency and decision-making. If de-biasing techniques are used when creating these AI systems, hiring can be less biased and allow for more diversity. However, there are instances where algorithms can be biased if not managed correctly. Automation can cause an automation paradox, and

managers often do not use AI-powered tools if they think it will be wrong more than it will be right (perception bias). Friction can occur between algorithmic decisions and human managers when there is a lack of alignment. Diversity can be determined by the organizations under which the AI operates, specific guidelines set for diversity, and ethical usage and design of the AI. Future studies should address the long-term effects of AI on recruitment. It is still unknown how this will affect diversity (considering all intersections of diversity) in the long term. There is not much information on how different de-biasing techniques can affect AI recruiting algorithms or how these algorithms may detect bias in real time and adjust. The effect of AI literacy on recruiters and hiring managers should also be researched to understand how it can affect their decision not to use AI when it is available. More research and the creation of standard ethical frameworks are needed. Finally, it should be determined whether emerging technologies such as generative AI and NLP should be included in AI recruiting software.

REFERENCES

1. Shneiderman B. Human-centered artificial intelligence: reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*. 2020;36(6):495-504. doi:10.1080/10447318.2020.1741118.
2. Haesevoets T, De Cremer D, Dierckx K, Van Hiel A. Human-machine collaboration in managerial decision making. *Computers in Human Behavior*. 2021;119:106730. doi:10.1016/j.chb.2021.106730.
3. Lacroux A, Martin-Lacroux C. Should I trust the artificial intelligence to recruit? Recruiters' perceptions and behavior when faced with algorithm-based recommendation systems during resume screening. *Frontiers in Psychology*. 2022;13:895997. doi:10.3389/fpsyg.2022.895997.
4. Awad E, Balafoutas L, Chen L, Ip E, Vecci J. Artificial intelligence and debiasing in hiring: impact on applicant quality and gender diversity. 2023. Available from: doi:10.2139/ssrn.4626059.
5. Will P, Krpan D, Lordan G. People versus machines: introducing the HIRE framework. *Artificial Intelligence Review*. 2023;56:1071-1100. doi:10.1007/s10462-022-10193-6.
6. Wilkens U, Lupp D, Langholf V. Configurations of human-centered AI at work: seven actor-structure engagements in organizations. *Frontiers in Artificial Intelligence*. 2023;6:1272159. doi:10.3389/frai.2023.1272159.
7. Bursell M, Roumbanis L. After the algorithms: a study of meta-algorithmic judgments and diversity in the hiring process at a large multisite company. *Big Data & Society*. 2024;11(1). doi:10.1177/20539517231221758.
8. Bano M, Zowghi D, Mourao F, Kaur S, Zhang T. Diversity and inclusion in AI for recruitment: lessons from industry workshop. *arXiv*. 2024.
9. Pant N, Joshi M. The impact of artificial intelligence on workforce diversity and inclusion: An HR perspective. *Educational Administration: Theory and Practice*. 2024;30(1):4503-12.
10. Zhang P. Application of artificial intelligence (AI) in recruitment and selection: the case of Company A and Company B. *Journal of Business and Management Studies*. 2024;6(3):224-225. doi:10.32996/jbms.2024.6.3.18.
11. Wilkens U, Lutzeyer I, Zheng C, Beser A, Prilla M. Augmenting diversity in hiring decisions with artificial intelligence tools. *The International Journal of Human Resource Management*. 2025;36(14):2585-2622. doi:10.1080/09585192.2025.2492867.