Bias in AI-Driven Candidate Selection Processes

Aleksandar Ćosić
University of Belgrade
Faculty of Organizational Sciences
Belgrade, Serbia
ac20245090@student.fon.bg.ac.rs
[0009-0002-1312-4013]

***Abstract* - This paper examines the nature, causes, and implications of bias in artificial intelligence (AI)-driven candidate selection processes, with a focus on the need for ethical development and application of AI systems. The primary objective is to identify and explain the key causes of bias in AI-driven candidate selection through a review and analysis of existing literature, contemporary studies, and case examples. The contribution of this work lies in systematic literature review and identification of the causes of the problem in the AI - driven candidate selection process.**

Keywords—bias, artificial intelligence, candidate selection

I. INTRODUCTION

The integration of artificial intelligence (AI) into employee selection processes has transformed recruitment by increasing efficiency and scalability. However, recent studies highlight significant concerns regarding inherent biases in AI systems, which may result in discriminatory hiring practices.

AI has become a crucial tool in candidate evaluation. The digital transformation of human resource management has led to the widespread adoption of algorithmic solutions in hiring processes. AI is used to analyze résumés, conduct predictive evaluations, and manage digital interviews [1].

The underlying idea of employing AI in recruitment is to establish higher standards that are independent of the attitudes and beliefs of individual recruiters [2]. AI enables companies to process a high volume of applications and ideally makes the recruitment process faster, more efficient, and less prone to human prejudice [3].

While these technologies enhance operational efficiency, an increasing number of research suggests that algorithms can replicate and even reinforce biases present in historical data [4].

This paper explores the nature of AI bias in employee selection, its root causes, and its consequences for both organizations and candidates. It also discusses strategies for mitigating such biases and ensuring the ethical application of AI in candidate selection.

II. THEORETICAL FRAMEWORK

The integration of artificial intelligence (AI) into recruitment processes has significantly transformed how organizations identify and select job candidates. AI-based systems are capable of processing large volumes of candidate data—including résumés, cover letters, and social media profiles—to detect patterns and assess candidate suitability for specific roles. These tools utilize machine learning algorithms to rank applicants, eliminate those who do not meet predefined criteria, and even predict future job performance based on historical hiring data [5]. Such automation can drastically reduce the time and costs associated with human-led hiring, potentially making the process more efficient, consistent, and data-driven [6].

AI bias in candidate selection refers to the systematic favoring or discrimination exhibited by AI systems during the selection process. This bias can manifest in various stages, including résumé screening, interview evaluation, and candidate ranking [7]. For example, it has been documented that some AI systems tend to favor résumés with names associated with white male candidates over those linked to black or female candidates, even when qualifications are identical [8].

Such bias is not necessarily the result of intentional discrimination but often stems from structured imbalances in the data, where algorithms "learn" to associate certain patterns with desirable outcomes without recognizing the socially problematic implications of those patterns [9]. For instance, if historical data indicates that interview scores were highest among candidates from a particular demographic group, the algorithm may internalize this pattern and continue applying it without contextual awareness.

III. METHODOLOGICAL FRAMEWORK

The primary approach employed in this study is a systematic literature review. This methodology was conducted in several stages (see Figure 1). In the initial phase, academic sources were collected and searched using online databases such as Google Scholar and arXiv, as well as through general web searches. The key search terms used included: AI employee selection, algorithmic bias, artificial intelligence and discrimination, and hiring discrimination.

From this broader collection, the literature was imported into the Mendeley reference manager, where duplicate entries were removed. Subsequently, abstracts were reviewed to determine the relevance of the studies to the topic of this paper, further narrowing down the list of sources used in the final analysis.

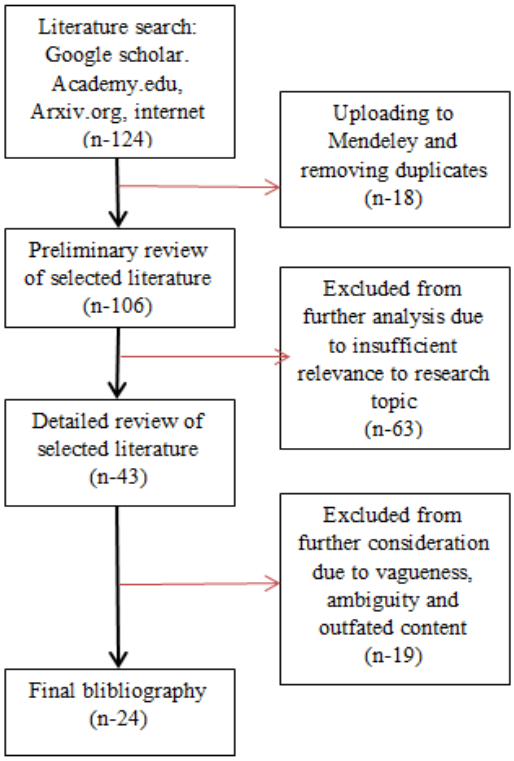

Figure 1 – Literature Review Process for Identifying Key Problem Areas

The selected literature was systematically analyzed to extract the most critical problem domains relevant to this study and to enable deeper investigation. This step in the research process revealed the following focal areas:

- Identification of types of bias in AI-driven candidate selection
- Identification and explanation of the causes of bias
- Consequences of bias
- Methods for mitigating the effects of bias in AI-based hiring systems

The literature analysis established a conceptual framework that provides a more detailed explanation of the challenges related to AI-guided personnel selection (see Figure 2).

For the purposes of this study, only literature published within the last 10 years was considered, as the topic addressed is relatively recent.

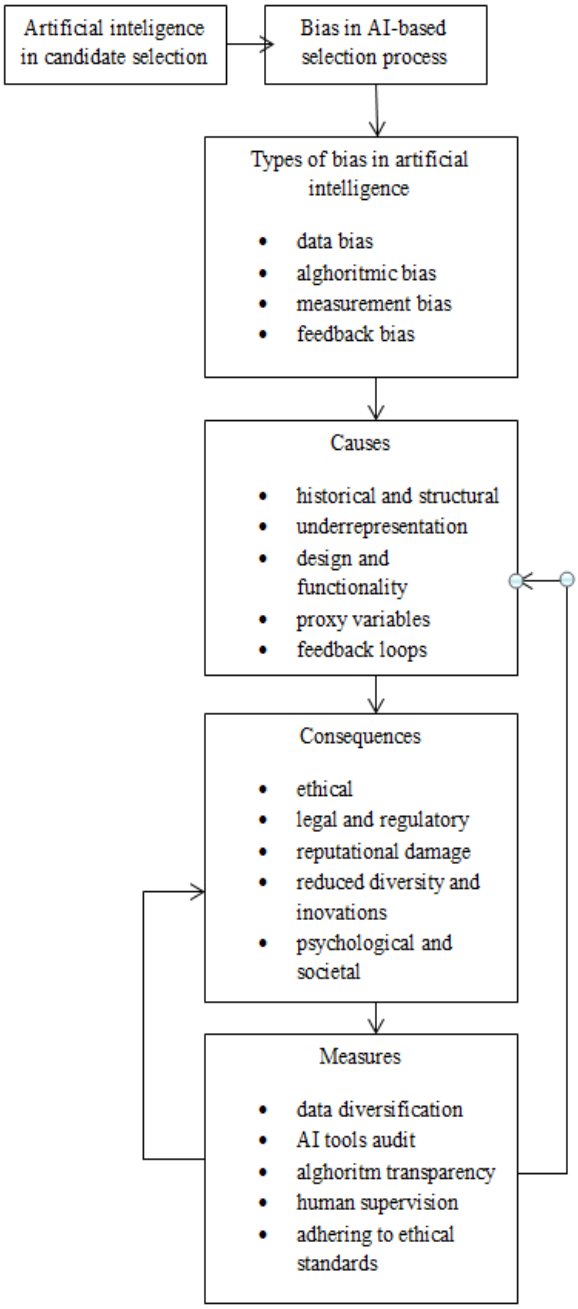

Figure 2 – Identified Problem Areas and Their Interdependence

IV. ANALYSIS OF RESULTS

The use of artificial intelligence (AI) in candidate selection processes is increasingly presented as an effective solution for enhancing human resources. However, it simultaneously introduces a range of ethical, technical, and societal challenges. One of the most serious issues is the emergence of bias at various stages of AI system development and application. This chapter analyzes specific types of bias that arise in the employment context, their causes, consequences, and potential mitigation strategies. The aim is to highlight the complexity of this phenomenon and emphasize the necessity of a comprehensive and responsible approach to implementing AI solutions in human resources.

A. *Types of Bias in Artificial Intelligence*

In the context of employment, AI bias can manifest across multiple layers of the process, from the data used to train models to the mechanisms through which algorithms make and learn from decisions. Understanding these types of bias is essential for developing fairer and more accountable systems. AI bias can generally be divided into several categories: data bias, algorithmic bias, measurement bias, and feedback bias [10].

• Data Bias

The data used to train AI systems often contain patterns of historical discrimination. If historical data reflect preferences for certain groups, the model may learn to favor them [11]. For instance, if hiring data from a company predominantly include male candidates, the algorithm might "learn" that being male is an implicit indicator of success, overlooking actual qualifications [12]. Similarly, datasets that ignore diversity in education may marginalize candidates from alternative educational systems.

• Algorithmic Bias

An algorithm that favors one group may perform less effectively for others [13]. The way a model is designed, including optimization metrics and model structure, can lead to biased outcomes even when data are not overtly discriminatory [4].

For example, if a model prioritizes metrics such as prediction accuracy or processing speed, it may neglect ethical considerations. Some models (e.g., neural networks) function as "black boxes," meaning users often lack insight into how the model reached a particular decision [14]. This lack of transparency prevents the identification of decision-making chain elements that led to undesirable or unethical outcomes, reducing the possibility for intervention and correction [15].

Additionally, how various input features are weighted can inadvertently favor or discriminate against certain candidate groups.

• Measurement Bias

Standard model accuracy metrics often do not incorporate fairness, potentially leading to systematic neglect of poor performance on data subsets [16]. For instance, cognitive ability tests that do not account for linguistic or cultural differences may yield unfavorable results for candidates with different educational or experiential backgrounds. Measurement bias can also arise in how AI systems interpret résumé content, especially if they rely on keyword frequency rather than contextual understanding. Such approaches favor candidates who have been trained to "optimize" their résumés for algorithms rather than those with the most relevant skills. In practice, this means that systems overestimate candidates using formal or technical language while underestimating those whose expressions deviate from standard formats, despite having equal qualifications [17].

The result is an evaluation that does not reflect the candidate's actual competence.

• Feedback Bias

Feedback bias refers to the tendency of algorithmic systems to confirm and reinforce patterns from prior decisions, particularly when using learning mechanisms such as reinforcement learning (RL) or online learning. In such cases, AI systems use their own past decisions as a basis for future predictions, potentially creating self-reinforcing loops of bias that intensify over time.

Systems learning from their own outputs (e.g., through reinforcement learning) may entrench discriminatory patterns if errors are not explicitly corrected [18]. For example, if the system consistently rates candidates from a specific demographic group as lower-performing, this assessment may be used as a valid signal in future decision cycles. Over time, this dynamic can produce a feedback loop in which biased patterns are not only maintained but amplified.

These types of bias rarely occur in isolation but rather in interaction. In practice, a combination of biased data, suboptimal algorithms, and poorly designed metrics can produce complex and difficult-to-detect forms of discrimination. Understanding these dimensions is crucial to developing ethically responsible AI solutions in human resources.

B. *Causes of Bias in Artificial Intelligence Used in Selection*

Bias in AI used for candidate selection does not stem solely from data deficiencies but results from a complex interplay of historical inequalities, technical limitations, and design choices in model development. Understanding these causes is key to building fairer and more accountable recruitment systems.

• Historical and Structural Bias in Data

A primary source of AI bias is the use of historical employment data that reflect existing societal inequalities. If AI models are trained on data from past practices that favored certain demographic groups, such as white males, they may learn to replicate those patterns. For instance, if a company has historically hired predominantly men for technical roles, a model trained on such data may rank female candidates lower, even when they possess the same qualifications [8].

• Underrepresentation of Diverse Groups in Data Classes

Bias also occurs when certain demographic groups are underrepresented in training datasets. For example, if a small proportion of résumés come from Black women, the AI may yield inaccurate assessments or exclude them entirely [19].

• Design and Functional Bias

AI systems are often optimized to maximize accuracy, efficiency, or predictive validity, metrics that may not align with fairness. For instance, if a model is rewarded for predicting job performance based on tenure or promotion history, and these outcomes are historically biased, the AI may unknowingly favor dominant groups [12].

• Proxy Variable Bias

Proxy variables statistically associated with protected characteristics like race, gender, or age can also introduce

bias. For example, zip codes, university names, or hobbies may serve as proxies for socioeconomic or racial background. Mujtaba and Mahapatra note that even seemingly neutral traits like writing style can act as indirect indicators of a candidate's race, gender, or education [19].

• Feedback Loops

One of the most complex and dangerous forms of algorithmic bias in hiring is the emergence of feedback loops. This occurs when an AI system continuously trains on its previous decisions and performance, thus consolidating and repeating historical patterns, including biases, through new decision-making cycles.

In practice, this means that if past recruitment cycles favored specific candidate profiles (e.g., men from technical universities), the system will classify these profiles as "successful" and automatically prioritize them in future evaluations, exponentially amplifying bias with each iteration.

C. Consequences of Bias in AI-Based Employee Selection

Algorithmic bias in candidate selection can lead to severe consequences for individuals, organizations, and society at large. Discriminatory outcomes threaten principles of fairness and equal opportunity [20].

Candidates unjustly rejected based on algorithmic assessments may lose trust in technology and institutions. In the long term, this may reduce workforce diversity, innovation, and damage organizational reputation [16].

Evaluating potential employees based on current employees perpetuates bias toward candidates who resemble those already hired [21].

• Ethical Consequences

From an ethical perspective, biased AI systems undermine the principle of fairness in hiring, candidates should be evaluated based on qualifications and merit, not demographics such as gender, race, or age. When AI systems reinforce existing societal inequalities, they contribute to discrimination and violate principles of equality and social justice.

Moreover, the use of non-transparent algorithms creates an ethical dilemma known as moral distancing, where responsibility for decisions shifts from humans to technology. This can decrease hiring managers' sense of accountability, potentially enabling unethical practices that would otherwise be challenged in human decision-making.

• Legal Risks and Regulatory Compliance

Legal frameworks in many countries explicitly prohibit discriminatory practices, even when unintentional or indirect. In the U.S., the Civil Rights Act prohibits employment discrimination based on race, gender, religion, or national origin. In the EU, the General Data Protection Regulation (GDPR) grants individuals the right to an explanation of automated decisions, including those related to hiring.

New regulations, such as New York's Local Law 144 and the EU AI Act, further address this issue.

In July 2023, New York became the first jurisdiction globally to mandate bias audits for commercial algorithmic systems, particularly those used in employment decisions. Local Law 144 (LL 144) requires annual independent audits for racial and gender bias and mandates public disclosure of audit reports. Employers must also provide transparency notices in job postings [22].

According to the EU AI Act, AI systems used in recruitment, candidate assessment, and hiring decisions are classified as "high-risk."[22] This includes:

- Automated résumé screening tools
- AI-based video interview and psychometric analysis
- Automated candidate ranking or scoring

Organizations using AI in hiring must ensure:

- Transparency: Candidates must be informed about AI usage and its influence on decision-making
- Human oversight: AI must not be the sole decision-maker; real human intervention is required
- Data quality and fairness: Training data must be representative to prevent bias
- Risk management: Employers must conduct risk assessments before using AI in selection
- Accountability: Documentation and audit trails must be maintained for compliance

Failure to comply with the EU AI Act may result in fines of up to €35 million or 7% of global turnover, depending on the severity of the violation.[22]

• Reputational Damage and Loss of Brand Trust

A company's reputation can be significantly harmed if it is revealed to use AI tools that discriminate against candidates. In the era of rapid information sharing via social media, a single case of malpractice can lead to widespread public backlash and consumer boycotts. Amazon's discontinuation of its AI recruitment tool, which was found to be biased against women, illustrates that even tech giants can suffer reputational damage. Companies seeking to maintain public trust and attract top talent must demonstrate responsibility in how they deploy technology.

• Reduced Workforce Diversity and Innovation

Diversity within teams is critical for fostering innovation, informed decision-making, and adaptability to changing markets. AI systems that systematically exclude candidates from marginalized groups produce a homogenous workforce where ideas tend to converge. According to Wilson and Caliskan, unchecked algorithmic bias can institutionalize uniform thinking, stifling creativity and diminishing teams' ability to approach complex problems from diverse perspectives [8].

• Psychological and Societal Impact

At the individual level, candidates rejected due to AI bias may experience a sense of injustice, helplessness, and reduced self-esteem. These effects can be particularly harmful to members of vulnerable groups who already face barriers in the labor market.

On a societal level, the continued use of biased AI systems can deepen existing inequalities, hinder social mobility, and erode public trust in technology and institutions. Over time, such systems may undermine efforts toward a more inclusive society and widen the gap between privileged and marginalized communities, pushing the latter further from the labor market.

D. Mitigating Bias in AI-Based Candidate Selection

Although bias in AI systems cannot be completely eliminated, it can be significantly mitigated through strategic interventions at various stages of development and deployment. The following are key measures to reduce the negative impacts of biased AI in candidate selection processes:

• Diversifying Training Data

One of the main sources of AI bias is the lack of diversity in training data. Training models on historical datasets that reflect social inequities (e.g., underrepresentation of minority groups) leads to the replication of these biases. Diversification involves including a wide range of demographic, cultural, and professional characteristics to ensure models better represent the real population. This includes balance in terms of gender, race, age, disability, education, and geographic origin.

• Regular Auditing of AI Tools

AI tools require continuous monitoring to identify and correct unintended consequences of their use. Regular audits allow systematic tracking of outputs, identification of discriminatory patterns, and corrective action. For example, New York City's Local Law 144 mandates employers using automated hiring tools to conduct annual independent audits to assess potential bias. Such legislation sets a precedent for institutionalizing audit practices worldwide.

• Algorithmic Transparency

Algorithmic transparency, also known as explainable AI (XAI), refers to the ability of systems to clearly communicate how and why a decision was made. This is crucial in candidate selection, as it allows employers, candidates, and regulators to understand the basis of decisions and identify potential systemic flaws or biases. Transparency enhances trust in the technology and strengthens accountability [12].

• Human supervision

Fully automating hiring decisions can lead to serious ethical and legal issues. While AI can effectively filter candidates, final decisions should be confirmed by human experts. Human judgment enables contextualization, empathy, and consideration of non-quantifiable factors. The practice of "human-in-the-loop" is becoming a critical component of responsible decision-making systems, ensuring that decisions are fair, correctable, and ethical, in addition to being accurate [23].

However, this also raises questions about the experience and accountability of the individuals overseeing AI systems and the creation of a potential sense of "false security."

• Adhering to Ethical Standards and Regulations

In addition to technical and procedural measures, organizations must ensure their AI systems comply with current legal frameworks and ethical guidelines. This includes adherence to principles of fairness, non-discrimination, privacy, and accountability. New legislative acts like the EU AI Act require strict compliance regarding risk classification and transparency. Ethics must not be an afterthought, but a foundational principle of AI development and application.

To ensure lawful AI use in employment under the EU AI Act, the following steps are essential:

- Conduct an impact assessment before deploying an AI system
- Provide candidates with the option to decline AI evaluation or request an alternative assessment method
- Implement mechanisms to detect and reduce bias
- Train HR staff to understand and monitor AI tools
- Collaborate with legal, regulatory, and data protection experts before and during AI implementation in selection

V. DISCUSSION AND CONCLUSION

Artificial intelligence (AI) holds the potential to revolutionize candidate selection processes by offering efficiency, consistency, and the ability to process large volumes of data. However, as numerous studies and real-world examples demonstrate, this potential comes with significant risks when AI is applied without critical oversight, ethical frameworks, and appropriate corrective mechanisms. Otherwise, instead of mitigating bias, AI may reinforce and exacerbate existing patterns of discrimination and social inequality.

Future research should focus on developing dynamic frameworks for bias auditing that can adapt to different employment contexts. This includes creating real-time fairness metrics and adaptive models that respond to fairness feedback.

These future research should address both technical and socio-ethical dimensions of AI bias in recruitment, encouraging interdisciplinary collaboration among computer scientists, ethicists, legal scholars, and HR practitioners.

Research Focus	**Description**
Dynamic Fairness Metrics	Develop adaptive fairness tools for diverse hiring contexts
Explainability in HR AI	Tailor XAI methods for HR professionals and candidates
Empirical Mitigation Studies	Long-term, real-world evaluations of bias reduction methods
Regulatory Impact Analysis	Study effects of laws like LL 144 and EU AI Act on hiring AI
Intersectionality & Bias	Examine compounded bias effects on intersecting identities
Feedback Loop Modeling	Analyze and break bias-reinforcing feedback cycles

Human-AI Collaboration	Optimize shared decision-making in recruitment systems
Ethical Embedding in AI	Integrate ethics upfront in AI design and development
Cultural & Linguistic Adaptation	Design inclusive AI for diverse cultural/linguistic groups
Candidate Psychological Impact	Investigate emotional effects of AI hiring rejections

Table 1. – Proposed future rearch fields with short description

Biases in algorithmic systems often originate from historical data, design choices, and inadequate model validation. If models are trained on data reflecting gender, racial, or class biases, there is a high risk that algorithms will not only perpetuate but amplify those biases through automated decision-making.

AI models—especially those based on deep learning, often function as "black boxes," making them difficult to interpret. More research is needed in developing explainable AI (XAI) tools tailored to HR domains, where explanations must be comprehensible to non-technical users such as managers and candidates. Candidates should have access to information about how they were assessed to improve transparency and trust. Recent studies have proposed methods for explainability in hiring systems, but further empirical testing is required [24].

As legal frameworks such as the EU AI Act and New York's Local Law 144 are increasingly enforced, future research should concentrate on developing AI tools that are ready for regulatory deployment. This includes documenting data sources, audit trails, and system behavior to meet transparency and accountability requirements. Ethical design principles—such as privacy by design and informed consent—must become integral to the AI lifecycle.

Because organizations deploying AI in recruitment bear significant responsibility, fair and accountable design must include the following components:

- Diverse and representative training data to reduce systemic bias
- Mandatory human supervision in decision-making to allow contextual and ethical evaluation of candidates
- Transparent and explainable models to ensure both candidates and authorities can understand the basis of decisions
- Compliance with legislative and ethical standards, including international regulations defining the limits of high-risk AI system deployment

In conclusion, the responsible use of AI in hiring must not be viewed solely as a technical challenge, but also as an ethical and legal imperative that requires cross-sector dialogue and ongoing reassessment of practices. Only through careful balancing of technological capabilities and social responsibility can we use AI to genuinely enhance fairness, inclusivity, and equality in the labor market.

REFERENCES

[1] K. Holstein, J. W. Vaughan, H. Daumé, M. Dudík, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," *Conf. Hum. Factors Comput. Syst. - Proc.*, 2019, doi: 10.1145/3290605.3300830.

[2] A. Miasato and F. Reis Silva, "Artificial Intelligence as an Instrument of Discrimination in Workforce Recruitment," *Acta Univ. Sapientiae Leg. Stud.*, vol. 8, no. 2, pp. 191–212, 2020, doi: 10.47745/ausleg.2019.8.2.04.

[3] Y. Acikgoz, K. H. Davison, M. Compagnone, and M. Laske, "Justice perceptions of artificial intelligence in selection," *Int. J. Sel. Assess.*, vol. 28, no. 4, pp. 399–416, 2020, doi: 10.1111/ijsa.12306.

[4] S. Hooker, "Moving beyond 'algorithmic bias is a data problem,'" *Patterns*, vol. 2, no. 4, p. 100241, 2021, doi: 10.1016/j.patter.2021.100241.

[5] P. van Esch, J. S. Black, and J. Ferolie, "Marketing AI recruitment: The next phase in job application and selection," *Comput. Human Behav.*, vol. 90, no. September 2018, pp. 215–222, 2019, doi: 10.1016/j.chb.2018.09.009.

[6] C. C. S. Liem *et al.*, *Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening*, no. November. 2018. doi: 10.1007/978-3-319-98131-4_9.

[7] Z. Chen, "Collaboration among recruiters and artificial intelligence: removing human prejudices in employment," *Cogn. Technol. Work*, vol. 25, no. 1, pp. 135–149, 2023, doi: 10.1007/s10111-022-00716-0.

[8] K. Wilson and A. Caliskan, "Gender, Race, and Intersectional Bias in Resume Screening via Language Model Retrieval," Jul. 2024, [Online]. Available: http://arxiv.org/abs/2407.20371

[9] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Comput. Surv.*, vol. 54, no. 6, 2021, doi: 10.1145/3457607.

[10] H. Suresh and J. Guttag, *A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle*, vol. 1, no. 1. Association for Computing Machinery, 2021. doi: 10.1145/3465416.3483305.

[11] E. Albaroudi, T. Mansouri, and A. Alameer, "A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring," Mar. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ai5010019.

[12] P. Seppälä and M. Małecka, "AI and discriminative decisions in recruitment: Challenging the core assumptions," *Big Data Soc.*, vol. 11, no. 1, Jan. 2024, doi: 10.1177/20539517241235872.

[13] Sheilla Njoto, "Gendered_Bots_Bias_in_the_use_of_Artificial inteligence in recruitment," 2020.

[14] K. Sokol and P. Flach, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," *FAT* 2020 - Proc. 2020 Conf. Fairness, Accountability, Transpar.*, pp. 56–67, 2020, doi: 10.1145/3351095.3372870.

[15] D. Shin and Y. J. Park, "Role of fairness, accountability, and transparency in algorithmic affordance," *Comput. Human Behav.*, vol. 98, no. November 2018, pp. 277–284, 2019, doi: 10.1016/j.chb.2019.04.019.

[16] M. Mitchell *et al.*, "Model cards for model reporting," *FAT* 2019 - Proc. 2019 Conf. Fairness, Accountability, Transpar.*, no. Figure 2, pp. 220–229, 2019, doi: 10.1145/3287560.3287596.

[17] A. Z. Jacobs, “Measurement and fairness,” *FAccT 2021 - Proc. 2021 ACM Conf. Fairness, Accountability, Transpar.*, pp. 375–385, 2021, doi: 10.1145/3442188.3445901.

[18] L. T. Liu, S. Dean, E. Rolf, M. Simchowitz, and M. Hardt, “Delayed impact of fair machine learning,” *IJCAI Int. Jt. Conf. Artif. Intell.*, vol. 2019-Augus, pp. 6196–6200, 2019, doi: 10.24963/ijcai.2019/862.

[19] D. F. Mujtaba and N. R. Mahapatra, “Fairness in AI-Driven Recruitment: Challenges, Metrics, Methods, and Future Directions,” May 2024, [Online]. Available: http://arxiv.org/abs/2405.19699

[20] I. D. Raji and J. Buolamwini, “Actionable Auditing Revisited: - Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products,” *Commun. ACM*, vol. 66, no. 1, pp. 101–108, 2022, doi: 10.1145/3571151.

[21] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, “Mitigating bias in algorithmic hiring: Evaluating claims and practices,” *FAT* 2020 - Proc. 2020 Conf. Fairness, Accountability, Transpar.*, pp. 469–481, 2020, doi: 10.1145/3351095.3372828.

[22] L. Wright *et al.*, “Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability,” *2024 ACM Conf. Fairness, Accountability, Transparency, FAccT 2024*, pp. 1701–1713, 2024, doi: 10.1145/3630106.3658998.

[23] B. Green and Y. Chen, “The principles and limits of algorithm-in-the-loop decision making,” *Proc. ACM Human-Computer Interact.*, vol. 3, no. CSCW, 2019, doi: 10.1145/3359152.

[24] Y. Rong *et al.*, “Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2104–2122, 2024, doi: 10.1109/TPAMI.2023.3331846.