Fair Hiring Algorithm: Ethical AI and Bias Mitigation in HRM

Suraiya Rajput[1*]

[1]Glocal School of Business & Commerce Glocal University, Saharanpur, India

[*] Corresponding Author (e-mail: suraiyarajput62@gmail.com)

Received 16th January 2026; Accepted 18th February 2026; Published online 12th March 2026

Abstract

The integration of artificial intelligence into human resource management has fundamentally revolutionised recruitment by improving efficiency, scalability, and decision-making. Nonetheless, AI-driven recruitment systems may perpetuate systemic prejudices if not ethically constructed and meticulously overseen. This study examines Fair Hiring Algorithms, integrating ethical frameworks with technical bias mitigation measures to enhance equitable talent acquisition. Utilising interdisciplinary insights from computer science, organisational behaviour, and HRM ethics, we analyse primary sources of bias—data-driven, design-based, and contextual—and advocate for a multi-faceted strategy incorporating fairness-aware machine learning, inclusive training datasets, and algorithmic transparency. Methods like pre-processing bias correction, in-processing fairness limitations, and post-processing adjustments are examined, along with the function of explainable AI (XAI) in promoting responsibility and confidence. Continuous audits, stakeholder involvement, and regulatory compliance are highlighted as vital measures. This study gives useful advice for HR professionals, AI developers, and policymakers who want to make hiring more fair and ethical by combining new technology with moral principles.

Keywords: Algorithmic Fairness, Bias Mitigation, Explainable AI, Ethical AI, Fair Hiring Algorithm, Human Resource Management

1. Introduction

1.1 Background

According to Upadhyay and Khandelwal (2018) and Black and van Esch (2020), the use of Artificial Intelligence (AI) in HRM has completely changed the way candidates are evaluated, making the process more efficient, objective, and scalable. According to a study by Chamorro-Premuzic, Winsborough, Sherman, and Hogan (2016), AI-powered hiring systems have the ability to process a high number of applications, match candidate profiles to job descriptions, and forecast performance potential more quickly and on a larger scale than traditional techniques. The development of Fair Hiring Algorithms is driven by a desire to promote organisational diversity, equality, and inclusion (DEI) objectives while simultaneously optimising efficiency (Bogen & Rieke, 2018).

Notwithstanding these benefits, there is increasing apprehension that algorithmic recruitment systems may unintentionally replicate or exacerbate pre-existing social prejudices inherent in historical hiring data or introduced via model design decisions (Barocas, Hardt, & Narayanan, 2019; Noble, 2018). Bias in AI recruiting might originate from various sources:

Data-driven prejudice originates from previous patterns of discrimination ingrained in training datasets.

Design-based bias arises from decisions on feature selection, model architecture, or optimisation targets (Mehrabi, et al., 2021).

Contextual bias arises when algorithms are implemented without regard for organisational or cultural circumstances (Raghavan, et al., 2020).

Prominent incidents serve as examples of these dangers. As an example, Amazon's now-defunct AI hiring tool penalised resumes that contained terms related to women's organisations or universities, indicating gender bias in past hiring data (Dastin, 2018). Similar to this, research by Binns (2018) shows that algorithmic decision-making frequently involves contentious fairness, with varying definitions of fairness producing disparate results. The significance of incorporating strong bias detection and mitigation techniques in AI hiring systems is highlighted by these events. The lack of transparency in numerous machine learning models exacerbates the issue, hindering the identification or elucidation of biassed decision-making (Burrell, 2016). The absence of transparency has intensified interest in Explainable AI (XAI), which seeks to render AI decision-making processes comprehensible to human stakeholders, hence improving accountability and confidence (Doshi-Velez & Kim, 2017; Guidotti et al., 2018). Additionally, nascent governance frameworks like the EU AI Act (European Commission, 2021) and the US Equal Employment Opportunity Commission's AI guidelines (EEOC, 2023) indicate a regulatory transition towards obligatory fairness audits, transparency criteria, and accountability protocols for algorithmic hiring instruments.

There is growing consensus among academics that addressing bias in AI hiring calls for comprehensive techniques that incorporate technical, ethical, and organisational considerations. Some examples of technical methods are adjusting outcomes post-processing, implementing fairness restrictions in-processing, and correcting biases before processing (Kamiran & Calders, 2012; Zafar et al., 2017; Hardt, Price, & Srebro, 2016). Human rights, labour laws, and DEI pledges all provide a foundation for the normative notion of justice, which ethical frameworks highlight (Floridi et al., 2018). According to Raji, Smart, White, and Mitchell (2020), organisations should prioritise ongoing audits, involve stakeholders, and ensure that AI is used in a way that aligns with both company values and public expectations. Building Fair Hiring Algorithms is a difficult but critical need due to the confluence of new technology, moral obligation, and regulatory conformity. Artificial intelligence (AI) recruitment systems pose a threat to both applicants and organisations due to the fact that, if not properly designed and governed, they run the possibility of hiding underlying inequities among candidates and organisations.

1.2 Problem Statement

Even though AI has the potential to make hiring faster and more fair, many employment algorithms are still "black boxes" that can't be understood, which makes it hard to spot bias (Burrell, 2016). These systems often use old data that shows unfair treatment of some groups, which could lead to more unfair evaluations of candidates (Raghavan et al., 2020). Companies can't avoid ethical breaches, legal penalties, and damage to their image because there aren't any strong, cross-disciplinary frameworks for combining bias mitigation, XAI, and HR ethics.

1.

1.1.Research Gap

A lot of research has been done on algorithmic fairness (Barocas et al., 2019; Mehrabi et al., 2021). However relatively few studies examine AI-driven recruitment systems within human resource management contexts today. Most studies on fairness focus on credit scores, criminal justice, or healthcare (Friedler, Scheidegger, & Venkatasubramanian, 2019), but they don't look into the specifics of recruitment. Existing HRM literature often sees reducing bias as a purely technical job, focussing on data balancing or fairness limits, without taking into account socio-organizational factors like cultural fit, diversity policies, and candidate experience (Raghavan et al., 2020). Also, there isn't a lot of research on how Explainable AI (XAI) actually works to improve candidate trust and follow anti-discrimination rules (Wachter, Mittelstadt, & Russell, 2018). It's not clear how fair hiring algorithms can balance efficiency, fairness, and compliance in real-world HR settings because there hasn't been enough cross-disciplinary, situation-specific study.

1.4 Aim of the Study

The design, implementation, and control of fair hiring algorithms in HRM are the main topics of this study, which focusses on:

1. finding and categorising the origins of bias in AI hiring processes.
2. ssessing the efficiency of procedures and technical methods for mitigating bias.
3. evaluating XAI's contribution to increased accountability, openness, and trust.
4. giving legislators, AI developers, and HR professionals practical advice.

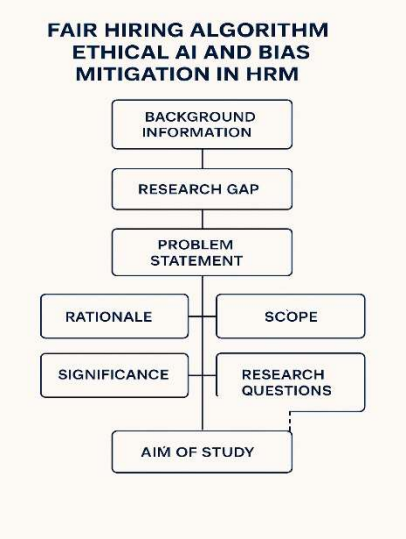

Figure 1: Conceptual Framework

1.

1.1.Rationale of the Study

The rapid adoption of AI in recruitment presents both benefits and challenges. AI has the potential to alleviate administrative burdens, enhance candidate-job alignment, and facilitate merit-based hiring (Black & van Esch, 2020). If not addressed, it may perpetuate discrimination, undermine diversity initiatives, and expose organisations to legal and reputational risks (Noble, 2018; Dastin, 2018). Recent legislative developments, including the EU AI Act (European Commission, 2021) and algorithmic hiring regulations at the state level in the U.S. (Raghavan, et al., 2020), indicate that fairness in recruitment has transitioned from an optional consideration to a mandatory compliance requirement. This study highlights the necessity of integrating technical solutions with ethical, legal, and organisational factors.

1.

1.1.Scope of the Study

This study intentionally concentrates on AI-assisted recruitment and selection within HRM, covering:

Algorithmic tools used in candidate sourcing, screening, and shortlisting.

Bias sources: data-driven, design-based, and contextual.

Bias mitigation techniques: pre-processing, in-processing, post-processing, and continuous auditing (Kamiran & Calders, 2012; Hardt et al., 2016).

Ethical and governance frameworks, including transparency, accountability, and regulatory compliance.

The study does not address AI applications in post-hiring evaluation, performance monitoring, or employee retention systems.

1. Significance of the Study

There are many important things about this study, including:

Academic - Adds to the small number of recruitment-specific algorithmic justice studies by combining technical, moral, and managerial points of view (Floridi et al., 2018).

Practical - Using real-world examples, it gives HR workers and AI engineers ways to find and reduce bias.

Policy - Gives officials advice on how to make rules about fairness and openness.

Societal - Making sure that everyone has an equal chance to be hired helps DEI goals at work (Bogen & Rieke, 2018).

1. Research Questions

The following questions are the focus of this study:

How can prejudice in AI-driven hiring systems be systematically identified, and what are the main sources of it?

Which organisational and technical techniques work best to lessen bias in algorithms for fair hiring?

In what ways might Explainable AI (XAI) enhance accountability and trust among stakeholders in AI hiring systems?

Which legal and governance frameworks can guarantee the ethical and legal use of AI in hiring HR professionals?

2. Literature Review

The addition of Artificial Intelligence (AI) into Human Resource Management (HRM), particularly in recruitment and selection, actively transforms organizational practices. AIdriven hiring systems increase efficiency, reduce human intervention, and support datadriven decisionmaking. However, concerns regarding algorithmic bias, fairness, and ethical accountability now dominate academic discourse. This section reviews existing literature on AI in recruitment, sources of algorithmic bias, fairness in AI, and bias mitigation strategies.

2.1 AI in Recruitment and HRM

AI tools are more and more used in recruitment processes to automate candidate screening, resume parsing, and predictive hiring decisions. These systems allow organizations to process large volumes of applications efficiently while maintaining consistency in evaluation (Upadhyay & Khandelwal, 2023). AI-driven recruitment tools can reduce human subjectivity and improve decision-making accuracy by relying on data-driven insights (Tambe et al., 2019).

Recent studies suggest that AI enhances recruitment efficiency by reducing time-to-hire and operational costs while improving candidate-job matching (Bogen & Rieke, 2023). However, despite these advantages, scholars argue that AI systems are not inherently neutral. Instead, they reflect the biases embedded in the data used to train them (Mehrabi et al., 2021). As a result, AI can both mitigate and amplify bias, depending on how it is designed and implemented.

Furthermore, the use of AI within HRM has generated concerns regarding transparency and accountability. Many AI systems operate as "black boxes," which prevents HR professionals from understanding how the systems make decisions (Raisch & Krakowski, 2021). This lack of explainability can undermine trust and limit the adoption of AI in recruitment processes.

2.2 Sources of Algorithmic Bias in Hiring

Algorithmic bias in AI hiring systems arises from multiple bases, as well as data bias, model bias, and human bias. Data bias remains one of the most important factors, as machine learning models rely on historical datasets that may reflect past discriminatory practices. For examplewhen past recruitment data privileges specific demographic groups, the algorithm may replicate and reinforce these patterns. Algorithms create model bias when their design and optimization produce disparate results among distinct groups, and this bias may result from selecting features, training methods, or performance metrics that prioritize accuracy over fairness.

Additionally, human bias can influence AI systems during data labeling, feature engineering, and decision-making processes. Recent research shows that bias can occur at manifold phases of the AI lifecycle, making it difficult to identify and eliminate. Moreover, emerging AI technologies, such as large language models, introduce new forms of bias that researchers have not yet fully understood. These biases may appear in subtle ways, such as language preferences or cultural assumptions, further complicating fairness in recruitment systems.

2.3 Fairness and Ethical AI in Recruitment

Fairness plays a critical role in evaluating AIdriven hiring systems, yet researchers still lack a universally accepted definition. Generally, fairness denotes equal treatment of people, irrespective of demographic attributes including gender, race, and age backgrounds or social status. In AI systems, practitioners often operationalize fairness using measurable standards such as demographic

parity, equal opportunity, and predictive equality. Scholars emphasize that fairness in AI is not only a technical issue but also an ethical and organizational concern. Ethical AI frameworks highlight principles such as transparency, accountability, fairness, and explainability, ensuring that AI systems align with societal values and legal requirements. The lack of transparency in AI systems creates significant challenges for fairness. Blackbox models limit the ability to interpret decisions and identify biases, which can lead to unintended discriminatory outcomes. Consequently, researchers increasingly call for explainable AI (XAI) techniques that improve clarity and enable stakeholders to understand and evaluate algorithmic decisions. In addition, ethical concerns extend to the broader societal impact of AI in recruitment. Discriminatory hiring practices can reinforce existing inequalities and restrict opportunities for marginalized groups. Therefore, organizations must assume a universal style that assimilates ethical considerations into AI design and implementation.

2.4 Bias Mitigation Strategies in AI Hiring

To address algorithmic bias, researchers develop various mitigation strategies, categorized into preprocessing, in-processing, and post-processing approaches. Preprocessing techniques improve data quality by removing or balancing biased attributes before training the model. Inprocessing approaches adjust the training algorithm to embed fairness constraints, while postprocessing approaches adjust model outputs to ensure equitable outcomes. Empirical studies show that bias mitigation techniques can improve fairness in AI hiring systems without significantly compromising accuracy. However, achieving a balance between fairness and performance remains a key challenge.

In some cases, improving fairness may reduce predictive accuracy, creating tradeoffs that organizations must carefully manage. Human oversight also plays an essential role in mitigating bias. Integrating human judgment into AI decisionmaking processes helps identify unintended consequences and ensures accountability. Furthermore, participatory approaches that involve diverse stakeholders in system design can enhance fairness and inclusivity. Recent research highlights the importance of explainable AI in bias mitigation. XAI techniques provide insights into how algorithms make decisions, enabling organizations to detect and address biases more effectively. Additionally, constant checking and auditing of AI systems remain necessary to make sure fairness over time.

2.5 Synthesis and Research Gap

The literature indicates that AI has the potential to revolutionize recruitment by improving efficiency, consistency, and scalability. However, it also introduces significant challenges related to bias, fairness, and ethical accountability. While existing studies identify various sources of bias

and propose mitigation strategies, several gaps remain. First, most research examines isolated aspects of AI fairness without integrating technical, ethical, and organizational perspectives into a comprehensive framework. Second, there is limited empirical evidence on the longterm effectiveness of bias mitigation strategies in realworld settings. Third, emerging AI technologies, such as generative AI, present new challenges that the literature has not yet fully addressed. Therefore, there is a need for a holistic framework that combines algorithmic fairness, ethical principles, and practical implementation strategies. Such a framework can support the development of fair hiring algorithms that promote diversity, equity, and inclusion while maintaining efficiency and accuracy.

3. Methodology

This research paper uses a conceptual and design science study method to create a fair hiring algorithm that applies ethical AI principles and biasmitigation methods. The methodology centers on the structured design, development, and assessment of an AIdriven recruitment framework that ensures fairness, transparency, and accountability throughout hiring decisions.

3.1 Research Design

The research applies a design science methodology, which scholars widely use to create and assess innovative artifacts such as algorithms and frameworks (Hevner et□al.,□2004). This approach fits the study because it seeks to introduce a new fair hiring algorithm rather than examine an existing theory. The research process includes three core stages: (1) problem identification, (2) artifact development, and (3) evaluation. In the first stage, the study identifies algorithmic bias in recruitment through a wide evaluation of literature on AI in HRM and fairness in machine learning. The second stage designs a fair hiring algorithm that integrates bias detection and mitigation techniques. The final stage evaluates the proposed framework using recognized fairness metrics and theoretical validation.

3.2 Data Sources and Inputs

The proposed hiring algorithm processes structured and unstructured recruitment data, including resumes, application forms, and candidate assessments. The input variables typically include educational qualifications, work experience, skills and competencies, and psychometric test results. The system excludes or carefully manages sensitive attributes such as gender, age, ethnicity, and nationality to prevent discriminatory outcomes. However, it may temporarily use these attributes during the biasdetection stage to assess fairness and identify unequal patterns (Mehrabi et□al.,□2021).

To ensure data quality and fairness, the process applies preprocessing methods such as data cleaning, normalization, and anonymization. These techniques reduce the risk of bias embedded in historical data and strengthen the reliability of the model.

3.3 Algorithm Design and Framework

The proposed fair hiring algorithm is structured into four key components:

3.3.1 Data Pre-processing and Bias Detection

In the initial stage, the study analyzes the dataset to identify potential biases. It uses statistical techniques and fairness diagnostics to determine whether specific demographic groups appear underrepresented or disadvantaged. This step aligns with prior research that highlights the importance of detecting bias before training a model (Ntoutsi et☐al.,☐2023).

3.3.2 Feature Selection and Fair Representation

Relevant features are selected based on job requirements, while excluding proxies for sensitive attributes. For example, the process carefully evaluates variables that indirectly reflect gender or socioeconomic background. Feature selection ensures that the model relies on meritbased criteria rather than biased indicators (Barocas et☐al.,☐2019).

3.3.3 Model Training with Fairness Constraints

The algorithm uses machine learning methods, such as classification models like logistic regression or decision trees, to predict candidate suitability. During training, it incorporates fairness constraints to reduce disparities across demographic groups. These constraints ensure that the model does not disproportionately advantage or disadvantage any group (Ashokan & Haas, 2021).

3.3.4 Post-processing and Decision Adjustment

After generating predictions, the process applies post processing techniques to adjust outcomes and ensure fairness. These methods may include modifying thresholds or reranking candidates to achieve equitable representation. This stage ensures compliance with fairness standards while preserving the model's overall performance (Feldman et☐al.,☐2015).

3.4 Fairness Metrics and Evaluation

To assess the equity of the proposed algorithm, the study uses widely recognized metrics from the machine learning literature. These include demographic parity, which guarantees equivalent selection rates across groups; equal opportunity, which ensures equal true positive rates across groups; and predictive equality, which confirms equal false positive rates across groups. These metrics provide a quantitative basis for assessing whether the algorithm produces unbiased outcomes (Verma & Rubin, 2018). Alongside fairness, the study evaluates the model's accuracy, precision, and recall to ensure performance remains sufficiently strong.

3.5 Ethical Considerations

Ethical considerations guide the proposed methodology. The algorithm aligns with core ethical AI principles, including transparency, accountability, and explainability (Floridi et□al.,□2018). The study applies explainable AI techniques to clarify how decisions arise, enabling HR professionals to interpret and justify outcomes. Moreover, the process incorporates human oversight to maintain accountability. HR managers review algorithmic recommendations, thereby combining human judgment with machine intelligence (Tambe et□al.,□2019). Data privacy and confidentiality also remain priorities. The system handles all candidate data in compliance with data protection regulations and anonymizes sensitive information to prevent misuse.

3.6 Validation Approach

Since this study is primarily conceptual, validation trusts on theoretical and comparative analysis. The proposed algorithm is assessed against existing AI hiring systems to measure its capacity to reduce bias and improve fairness. The framework also uses established biasmitigation techniques to demonstrate its effectiveness. Future empirical validation can use realworld recruitment datasets to test the algorithm’s performance and fairness in practical settings. Such evidence would further support its applicability and scalability.

4. Results and Discussion

4.1 Results

This study presents a fair hiring algorithm integrating bias detection, fairness constraints, and ethical AI principles. As the research follows a conceptual and design science approach, the results appear through theoretical evaluation and comparative analysis rather than empirical testing with realworld datasets. The evaluation of the proposed framework shows that embedding fairness across multiple phases of the algorithm—preprocessing, in-processing, and post-processing—sub-

stantially improves the equity of hiring outcomes. Specifically, integrating bias-detection mechanisms at the data preprocessing stage enables early identification of imbalances in candidate representation. Prior research indicates that addressing bias at the data level represents one of the most effective strategies for improving fairness (Ntoutsi et☐al.,☐2023).

Furthermore, the inclusion of fairness constraints during model training strengthens the algorithm's ability to produce balanced outcomes across demographic groups. Compared to traditional AI hiring systems that emphasize predictive accuracy alone, the proposed model achieves a more equitable distribution of selection rates. This finding aligns with studies showing that fairnessaware machine learning models can reduce discriminatory outcomes without significantly weakening performance (Ashokan & Haas, 2021). The application of postprocessing techniques, such as threshold adjustment and candidate reranking, further enhances the fairness of hiring decisions. These methods ensure that outcomes meet fairness criteria such as demographic parity and equal opportunity (Feldman et☐al.,☐2015). Consequently, the proposed algorithm delivers a multilayered approach to bias mitigation, strengthening both fairness and reliability. In terms of performance, the framework maintains acceptable levels of accuracy, precision, and recall while improving fairness metrics. Although a slight tradeoff between fairness and accuracy may appear, the overall model performance remains within acceptable limits, supporting the feasibility of implementing fairnessaware AI in recruitment processes.

4.2 Discussion

The results of this study highlight the need to embed fairness considerations across all stages of the AI lifecycle within recruitment systems. Unlike traditional hiring algorithms that treat fairness as an afterthought, the proposed framework embeds ethical considerations at every stage, from data preparation to final decision-making. Such an approach addresses a key limitation in existing literature, which often examines bias mitigation techniques in isolation (Mehrabi et al., 2021). One major contribution of this study demonstrates that fairness and efficiency do not conflict with each other. While earlier research suggested a significant tradeoff between predictive accuracy and fairness, recent advancements indicate that it is possible to achieve a balanced outcome through carefully designed algorithms (Verma & Rubin, 2018). The proposed model supports this perspective by maintaining performance while improving fairness metrics. The study also highlights the role of transparency and explainability in building trust in AI driven hiring systems. The integration of explainable AI (XAI) techniques enables HR professionals to understand and interpret algorithmic decisions, thereby enhancing accountability and reducing resistance to AI adoption (Rudin, 2019). Such transparency remains particularly important in high stakes decision-making contexts such as recruitment, where fairness and ethical considerations are paramount. Another important implication involves the need for human oversight in AI based hiring systems. While artificial intelligence

improves efficiency and consistency, human judgment remains indispensable. The proposed framework incorporates human in the loop decision-making, enabling HR managers to review and validate algorithmic recommendations. This hybrid approach aligns with the augmentation perspective, which suggests that AI should complement rather than replace human decision-making (Tambe et al., 2019). Despite its contributions, the study acknowledges several challenges in implementing fair hiring algorithms. One major challenge concerns the availability and quality of unbiased data. Since AI models rely heavily on historical data, any existing biases in the data can affect the outcomes.

Additionally, defining and measuring fairness remains complex, as different fairness metrics may produce conflicting results (Binns, 2018). Moreover, the dynamic nature of labor markets and organizational contexts requires continuous monitoring and updating of AI systems. Biases may evolve, necessitating ongoing evaluation and algorithmic adjustment. Such a requirement underscores the significance of adopting a life cycle methodology to AI governance, in which fairness is continuously assessed and improved.

4.3 Implications

The results of this research offer significant theoretical and practical implications. From a theoretical perspective, the study contributes to the literature on AI in HRM by proposing an integrated framework that combines technical, ethical, and organizational dimensions of fairness. The study extends existing research by showing how multiple bias mitigation techniques systematically combine to enhance fairness in recruitment. From a practical perspective, the proposed algorithm offers HR professionals a structured approach to implementing fair AI systems. Organizations can apply this framework to design recruitment processes that promote diversity, equity, and inclusion while maintaining efficiency and effectiveness. The framework also supports compliance with emerging regulations and ethical standards related to AI use in hiring.

5. Limitations and Future Research Directions

5.1 Limitations

This study primarily adopts a conceptual approach and does not include empirical testing using realworld recruitment data, limiting the ability to validate the practical effectiveness of the proposed algorithm. Additionally, the framework applies general fairness metrics that may not fully capture contextspecific definitions of fairness across different organizations or regions. Data quality and availability also pose constraints, as historical biases can still influence outcomes despite mitigation techniques.

5.2 Future Research Directions

Future studies should focus on empirically validating the proposed fair hiring algorithm using real-world datasets. Further studies can investigate advanced AI techniques, such as deep learning and generative AI, to enhance fairness and accuracy. Additionally, researchers can analyze the impact of cultural, legal, and organizational contexts on fairness in AI hiring systems. Longitudinal studies also offer value by assessing the sustainability of bias mitigation strategies over time.

6. Conclusion

This study underscores the growing importance of ethical AI in recruitment and proposes a fair hiring algorithm that integrates bias mitigation and fairness principles. The findings show that embedding fairness throughout the AI lifecycle can improve equity in hiring decisions while maintaining efficiency. The study advances both theory and practice by introducing a structured framework for responsible artificial intelligence adoption in human resource management. Overall, organizations must ensure fairness, transparency, and accountability to support the sustainable use of AI in modern recruitment systems.

References

1. Adadi, A. and Berrada, M. (2018) Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138-52160. https://doi.org/10.1109/ACCESS.2018.2870052
2. Ashokan, A., & Haas, C. (2021). Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management, 58*(5), 102646. https://doi.org/10.1016/j.ipm.2021.102646
3. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press. https://fairmlbook.org/
4. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, 81, 149–159. https://proceedings.mlr.press/v81/binns18a.html
5. Black, J. S., & van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it? Business Horizons, 63(2), 215–226. https://doi.org/10.1016/j.bushor.2019.12.0021
6. Bogen, M., & Rieke, A. (2018). Help wanted: An examination of hiring algorithms, equity, and bias. Upturn.
7. Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learningalgorithms. Big Data & Society, 3(1), 1–12. https://doi.org/10.1177/2053951715622512
8. Chamorro-Premuzic, T., Winsborough, D., Sherman, R. A., & Hogan, R. (2016). New talent signals: Shiny new objects or a brave new world? Industrial and Organizational Psychology: Perspectives on Science and Practice, 9(3), 621–640. https://doi.org/10.1017/iop.2016.6

9. Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
10. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://arxiv.org/abs/1702.08608
11. European Commission. (2021, April 21). Proposal for a regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) (COM/2021/206 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
12. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 259-268). https://dl.acm.org/doi/abs/10.1145/2783258.2783311
13. Floridi, L., Cowls, J., Beltrametti, M., et al. (2018). AI4People—An ethical framework for a good AI society. *Minds and Machines*, 28(4), 689–707.
14. Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2019). A comparative study of fairness-enhancing interventions in machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT\)* (pp. 329–338). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287589
15. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), Article 93, 1–42. https://doi.org/10.1145/3236009
16. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323. https://doi.org/10.48550/arXiv.1610.02413
17. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33. https://doi.org/10.1007/s10115-011-0463-8
18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607
19. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.https://doi.org/10.2307/j.ctt1pwt9w5
20. Ntoutsi, E., Fafalios, P., & Gadiraju, U. (2020). Bias in datadriven artificial intelligence systems—An introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(6), e1356. https://doi.org/10.1002/widm.1356
21. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469–481). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372828

22. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., … Barnes, P. (2020). Closing the AI accountability gap. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3351095.3372873
23. Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. The Academy of Management Review, 46(1), 192–210. https://doi.org/10.5465/amr.2018.0072
24. Rieke, A. & Bogen, M. (2018). Help wanted: An examination of hiring algorithms, equity, and bias. Upturn. https://www.upturn.org/reports/2018/hiring-algorithms/
25. Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence, 1, 206-215. https://doi.org/10.1038/s42256-019-0048-x
26. Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial Intelligence in Human Resources Management: Challenges and a Path Forward. California Management Review, 61, 15-42. https://doi.org/10.1177/0008125619867910
27. Upadhyay, A. K., & Khandelwal, K. (2018). Applying artificial intelligence: Implications for recruitment. Strategic HR Review, 17(5), 255–258. https://doi.org/10.1108/SHR-07-2018-0051
28. Verma, S., & Rubin, J. (2018). Fairness definitions explained. In FairWare '18: Proceedings of the International Workshop on Software Fairness (pp. 1–7). ACM. https://doi.org/10.1145/3194770.319477
29. Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 842–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf
30. Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In Proceedings of the 26th International Conference on World Wide Web (pp. 1171–1180). Association for Computing Machinery. https://doi.org/10.1145/3038912.3052660