Journal of Commercial Law Review
2025, VOL. 9, NO. 3, 54-59
DOI: 10.47297/wspjclrWSP2516-249711.20250903

Toward a Fairer Future: Understanding and Addressing Algorithmic Bias in Hiring

Rui Zhong

Facul Kong Shatin, NewTerritories, HongKong, 999077, China

ABSTRACT
AI-driven hiring systems have gradually made a difference instead of traditional hiring patterns. However, such hiring algorithms with a mask of neutrality may lead to algorithmic bias, damaging equal employment rights, which is defined as the outputs of AI algorithms benefit or disadvantage certain individuals or groups more than others without justifications for such unequal impacts. The bias can stem from two stages of algorithm design including data collection and modelling. In the stage of data collection, historical data collected containing human bias and reflecting societal inequity will perpetuate the bias, and non-representative data collected violating diversity or completeness of data will also result in the bias. In the stage of modelling, irrelevant or redundant features selected like sensitive attributes and proxy attributes closely related to sensitive attributes will introduce the bias, and opaque black-box system with less explainability may contain unknown bias. Technically, the bias can be mitigated through debiasing datasets involved in diversifying data and reweighting and resampling in preprocess, introducing fairness metrics in modelling, and increasing transparency. Regarding with legal governance, the framework of AI fairness reporting and auditing should be constructed to identify bias and achieve transparency and liability of bias should be regulated to clarify who shall be liable and why shall be liable.

KEYWORDS
Algorithmic hiring; Algorithmic bias; Fairness

1 Introduction

Whether we are aware of it or not, artificial intelligence (AI) algorithms influence everything from the news we read to the jobs we get (Rakivnenko, 2024). In particular, companies are increasingly relying on AI-driven hiring systems replacing traditional hiring patterns depending on humans. In a late-2023 IBM survey of more than 8,500 global IT professionals, it's noted that 42% of companies were using AI screening to improve recruiting and human resources and another 40% of respondents were considering integrating the technology (Lytton, 2024). Why this trend happens? Superficially, it's about efficiency and speed. Through the use of AI algorithms, large volumes of data can be processed more quickly and the whole hiring process can work more smoothly including screening resumes, matching candidates with job requirements, and scheduling interviews by AI-powered tools. However, the essential reason should be considered more that it seems that AI algorithms are generally deemed as neutral which can end biases in the hiring process dominant by human beings, which can be explained from two aspects. First is mathematical foundation that decisions by AI are based on programmed rules and patterns and its operation has no personal emotions. Second is consistency in processing that information is processed in a standardized way, applying the same criteria uniformly without gut feelings.

Honestly, are AI algorithms neutral? The real and complicated world tells us the answer pitifully that it's NO. Therefore, the terminology "Algorithmic Bias" appears. Before analyzing why algorithmic bias exist, it's significant to realize what it means. Obviously, there are various definitions of algorithmic bias, or its opposite (algorithmic fairness) given by scholars and organizations. For example, algorithmic bias means that an assessment systematically over- or under-determines scores for a particular group (American Psychological Association, 2018). Algorithmic fairness should guarantee equal and just distribution of benefits and costs while ensuring that each individual and group could benefit from equal opportunities (The High-Level Expert Group on Artificial Intelligence, 2018). While there's agreement that fairness should address power asymmetries and guarantee dignity, autonomy, and equality, the translation into practical requirements varies based on stakeholders' expectations, specific context and competing interests (Rigotti & Fosch-Villaronga, 2024). From my perspective, algorithmic bias is defined as the outputs of AI algorithms benefit or disadvantage certain individuals or groups more than others without justifications for such unequal impacts, and such impacts are generally involved in factors such as disability, color, race, religion, gender, age and national origin.

When companies emphasize that their AI tools don't explicitly consider factors stated above leading to unequal impacts which however isn't sufficient to prevent bias in a way, the belief in algorithmic neutrality might make stakeholders less likely to question or verify the system's outputs. Such a belief that algorithms better overcome bias than their human counterpart will lead to a paradox of neutrality because this belief may lead to the opposite effect, as it could increase the possibility of missing or overlooking potential or actual bias for stakeholders (Marieke & Sach, 2024). Therefore, algorithmic bias hidden behind AI looks like scarier than superficial human bias in a hiring process. As Hilke Schellman (2024) said, one biased human hiring manager can harm a lot of people in a year, and that's not great. But an

algorithm that is maybe used in all incoming applications at a large company… that could harm hundreds of thousands of applicants. Since equal employment rights have been recognized internationally and one of its cores focuses on hiring decisions should be based on merit, skills, and qualifications rather than irrelevant personal characteristics, algorithmic bias is threating such fundamental human rights and has gotten much attention from the world.

2 Sources of bias

An algorithm-driven decision-making process begins with the collection of raw data from various sources. Once collected, the data undergoes preprocessing to prepare it for analysis. Relevant attributes, or features, are then selected to train the algorithm. By identifying key patterns in the data, machine learning models learn to generate outcomes based on these features. Once the algorithm is trained, it is used to make decisions or recommendations. Generally, the bias can be produced from two stages including data collection and modelling.

2.1 The stage of data collection

As the first stage in designing an algorithm, data collection involves gathering the information the algorithm needs to learn and work properly. This data can come from different sources such as databases, surveys, sensors, or user interactions. Once collected, the data is cleaned and organized to prepare it for the next stages of the process. In the stage of data collection, the formative logic of bias can be described as "garbage in, garbage out" (Houser, 2019). The data with nature of bias is input thus the consequent output is also biased influenced and when the biased algorithm output is fed back into the system, it can further reinforce bias which is called negative feedback loops creating a cycle where the algorithm learns and amplifies discriminatory patterns (Kordzadeh & Ghasemaghaei, 2021). Bias produced in the stage of data collection can divided into two different kinds, namely, the bias from historical data and the bias from non-representative data.

2.1.1 Bias form historical data

Historical data, capturing patterns of human behavior, institutional practices, and public norms in the past, may reflect the inequities and stereotypes embedded in society to favor certain specific demographics. Bias form historical data occurs when AI systems are trained on past data that contains human bias, causing the AI to learn and reproduce such bias in its decisions, which is particularly problematic in recruitment as it can perpetuate long-standing social inequalities and discrimination (Rigotti & Fosch-Villaronga, 2024). The algorithm is like a mirror to stronger reflect the biased history. In Amazon's case, Amazon developed an AI-based screening tool that ranked potential hires based on patterns observed in resumes from the past 10 years and the training data came predominantly from tech industry hiring processes which historically favored men (especially for technical roles such as software engineers or developers). The algorithm learned to favor traits and keywords commonly associated with male candidates and finally the AI tool was scrapped due to the gender bias revealed.

2.1.2 Bias from non-representative data

Unlike the bias from historical data rooted in actual inequities embedded in society, the bias from non-representative data relates to diversity or completeness of the data collected. The data used to train an algorithm is not representative of the broader population it's intended to serve (which also can describe as dataset imbalance) resulting in bias. Specifically, certain groups are underrepresented, excluded, or overrepresented in the training data collected and the model will overfit to the dominant majority group and fail to generalize well to minority groups. As a result, the AI may learn to favor candidates with the characteristics and experiences of the over-represented group while overlooking those belonging to underrepresented or excluded groups, which leads to biased decisions (Rafi, 2024). In a lawsuit filed against Workday, a company providing HR software, its AI hiring algorithm was claimed to show bias against black, older, and disabled applicants who faced lower hiring rates. The model was trained on demographically imbalanced datasets which likely overrepresented white, younger, and able-bodied candidates, to favored traits and qualifications more commonly found in dominant groups.

2.2 The stage of modelling

The modeling stage is a key step in designing an algorithm, focusing on training and evaluating a machine learning model based on the collected data to learn patterns and make decisions aligned with its intended purpose. The modeling process begins with training, where the model learns from the training dataset by identifying relationships between

inputs (features) and outputs (targets) and is optimized by mathematical techniques, such as minimizing errors or maximizing accuracy and validation data is then used to fine-tune the model's hyperparameters, ensuring it performs well on unseen data. Next, the model is evaluated on a test dataset to measure its real-world performance through metrics like accuracy, precision, recall, or mean squared error. In the stage of modelling, the bias is from two aspects including feature selection and proxy, and black box.

2.2.1 Bias from feature selection and proxy

Feature selection involves identifying the most relevant and significant features in a dataset. By focusing on the most useful features, feature selection helps the model better understand relationships in the data and makes it more efficient. How features are selected and weighted in the algorithmic decision-making process depends on specific choices made by algorithm designers in determining which attributes to include and prioritize (Wang, Wu, Ji, & Fu, 2024). Bias occurs when the process of feature selection introduces errors or distortions into the model negatively impacting its performance and accuracy due to irrelevant or redundant features are included or critical features are excluded.

Typically, bias results from sensitive attributes such as race, gender, age and religion which are included in the feature selection process to enable the model to make discriminatory decisions intentionally. For example, selecting gender as a feature might result in male candidates being favored for traditionally male-dominated roles, perpetuating workplace discrimination. However, if such sensitive attributes have been avoided, bias may also arise indirectly which is known as proxy bias. Through the use of proxy or masked attributes that serve as stand-ins for sensitive attributes, algorithms can be allowed to engage in "redlining" or other forms of discrimination while appearing to be neutral and objective (Gillis & Spiess, 2019). For example, algorithms may rank candidates based on their location, preferring applicants from affluent or urban areas and during this process, ZIP codes can act as proxies for race bias due to historic long-standing patterns of segregation. Algorithms may also give preference to candidates' names that sound American or familiar where a name can act as a proxy for ethnicity or religion bias because names from America, Africa, Asia, and Middle East have their own ethnical or religious characteristics.

2.2.2 Bias from black box

A black-box model refers to a machine learning model that operates as an opaque system where the internal workings of the model are not easily accessible or interpretable or the decision-making process and reasoning are not transparent to the user in other words (Hassija, Chamola, Mahapatra, Singal, Goel, Huang, Scardapane, Spinelli, Mahmud, Hussain, 2023). If AI outcomes cannot be explained or transparent, they may contain unknown bias. For algorithms with black box problem, lack of transparency or explainability makes it difficult to understand how they arrive at specific decisions.

There are two main reasons why black box exists. The first is model complexity. It's noted that advanced AI systems like deep learning have millions of parameters and layers, making it inherently strenuous to interpret how specific features influence final decisions. For instance, sometimes it's unclear why a neural network might reject one job applicant but approve another with nearly identical qualifications. Another is hidden interactions. Black box models often rely on complex and nonlinear interactions between features that are too intricate to trace, making it challenging to detect unintended correlations or discriminatory patterns. It's scary that unintentional employers blindly trust black box systems without awareness or scrutiny of the potential bias encoded in their logic hiding behind technical limitations which may reinforce societal discrimination. The lack of transparency or explainability, eroding trust in AI systems, prevents affected individuals or groups from contesting unfair decisions, regulators from enforcing accountability, and developers from correcting harmful outcomes.

3 Technical mitigation for bias

Turing award winner Donald Knuth said that computers ''do exactly what they are told, no more and no less.'' An algorithm can fulfill an objective in many ways, while still violating the spirit of said objective (Hooker, 2021). The consequent bias is firstly considered as a technical issue thus coming up with some technical ideas or taking some technical measures to mitigate such bias should be deemed necessary although the sole technical help is of insufficiency. Specifically, technical considerations mentioned below are mainly designed for employers to mitigate the bias revealed in their hiring algorithms. There are three main focuses including debiasing datasets, introducing fairness metrics in modelling, and increasing transparency of algorithms.

3.1 Debiasing dataset

Diversity of training data is crucial which involves ensuring that examples from all demographic groups are adequately represented. For instance, in a facial recognition system, if the training dataset just mainly includes faces of lighter-skinned individuals, the algorithm may perform poorly on darker-skinned individuals. For mitigation, additional data from underrepresented groups can be collected or publicly available datasets (like inclusive benchmarks such as PULSE or FairFace) can be added to focus on balanced representation of diverse groups. Data augmentation techniques can also be considered in some situations (Albaroudi, Mansouri, & Alameer, 2024).

In the preprocessing for data, reweighting and resampling will help to mitigate bias. Reweighting is to assign higher weights to the underrepresented groups in the dataset so that their examples are treated as more important during training. For instance, if 80% of training samples belong to one group and only 20% to another, this imbalance can be accounted for by assigning proportional weights (like 1.25 for the minority group and 0.80 for the majority group) which enables the model to focus more on learning patterns for the minority group. As for resampling, regarding with imbalanced datasets, oversampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) can be used to duplicate data points from underrepresented groups or undersampling to reduce the number of examples from overrepresented groups.

3.2 Introducing fairness metrics in modelling

Through the introduction of fairness metrics into algorithmics, they can become more fairness-aware to mitigate the bias. The first step is to define measurable fairness goals that align with the desired outcomes. Fairness metrics provide a way to quantify and evaluate the bias in an algorithm, beneficial to monitor and address unequitable treatment of different groups. There are some common fairness metrics including demographic parity ensuring equal positive outcomes across groups and equal opportunity ensuring that individuals who qualify for a favorable outcome have an equal chance of being selected across all groups. Based on the context of the actual use, the appropriate metric selected can enable employers systematically assess whether the algorithm is treating every group fairly.

Once fairness metrics are identified, they can be integrated into the development and evaluation of the algorithm. During training, fairness constraints can be added to the model, guiding it to optimize both its accuracy and fairness goals. For example, if a model recommends consistently candidates from one demographic group over others, fairness constraints will adjust the model to balance its selections while still focusing on qualified candidates. Fairness metrics also serve as evaluation tools. After training, the outcomes of the model should be analyzed to examine that it meets fairness thresholds such as consistent error rates or balanced decisions for minority and majority groups. If the outcomes can't achieve fairness goals, further adjustments can be taken by modifying the training data, redefining decision thresholds, or retraining the model with additional safeguards, which enable fairness metrics to act as a continuous checkpoint helping to mitigate bias and ensure equitable outcomes across all groups (Albaroudi, Mansouri, & Alameer, 2024).

3.3 Increasing transparency of algorithms

To reduce the risk of bias caused by deviant algorithms involved in the black-box dilemma, increasing transparency can facilitate remediation as it allows for the detection, explanation, and correction of unfair outcomes (Chen, 2023), which requires a more understandable decision-making process of algorithms. Explainable AI (XAI) can be considered to clarify how a model arrives at its decisions. For example, XAI approaches like LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations) help explain outcomes by showing the contribution of each input variable to the outcome. LIME works by approximating the model locally around a specific decision, making it easier to understand why the model behaved in a particular way for a specific instance. SHAP, on the other hand, assigns importance weights to each feature, quantifying their individual impact on the decision. These tools empower stakeholders to understand algorithmic decisions in detail and identify whether the model is relying on fair and appropriate factors.

4 Legal governance for bias

Algorithmic bias in hiring is related to not only technical considerations, but also legal governance. There are two main focuses including AI fairness reporting and auditing and liability for bias. The fairness reporting framework is beneficial to identify the bias and achieve the goal of transparency, and liability for bias is dealing with issues that who shall be liable and why shall be liable for such bias.

4.1 AI Fairness reporting

The framework of AI fairness reporting has positiveness both for candidates and employers. For candidates, the framework helps identify and further mitigate bias through achieving transparency to gives candidates confidence that the AI system has been vetted for equity and non-discrimination and to clarify employers' accountability for potential harm by bias. For employers, the framework reduces reputational and legal risks because algorithmic bias may lead to significant lawsuits and penalties, damaging trust from the labor market.

An ideal thinking of AI fair reporting obligations for employers can be based on disclosure of models, interventions and training datasets (Yap & Lim, 2022). Employers must disclose all machine learning models used for hiring and models not directly affecting individuals or groups should also be disclosed because they might raise some fairness concerns (like unfair sampling). Also, employers must disclose any interventions they made to reveal potential implications of the chosen debiasing procedures whether intentionally or unintentionally. Additionally, employers must release all datasets used for training models for public inspection or third-party audit. Any use of open-source datasets and pre-trained models from third parties must be disclosed so that it can be considered that whether known biases in open-source datasets may affect the employer's AI models. The release of datasets enables independent verification of fairness claims. However, sometimes it's challenging to implement such requirements in reality due to reasons such as protection of trade secrets, privacy of data and compliance costs. If such requirements can't be implemented, employers must give justifications to explain satisfactorily for non-compliance which reflects the flexibility of framework.

4.2 Liability for bias

When algorithmic bias occurs unfortunately, the most crucial thing is to determine legal liability for such bias, which depends on multiple factors based on different jurisdictions. However, the common logic can be general in a way, focusing on who shall be liable and why shall be liable. Two legal models in US and EU will be revealed, and they are now representative in algorithmic liability patterns in the world. Interestingly, US's attitude is reactive, with litigation-driven relief afterwards, where AI-focused laws' creation is slower than AI's development, and reversely, EU's attitude is proactive, with prior preventative regulation, where AI-targeted laws are more advanced than AI's progress.

In US, liability for algorithmic bias is distributed based on sector-specific anti-discrimination statutes and general tort principles, without AI-specific laws, on a case-by-case basis. Liability often falls on deployers using algorithms in decision-making because they apply AI tools to real-life situations, where the bias manifests and causes harm. The doctrine of disparate impact allows deployers liable even for unintentional harm if the algorithmic outcomes disproportionately affect one group more than another. However, deployers can defend the biased practice if they can prove that it's justified by business necessity and no less biased alternatives can be available. Sometimes, AI developers can be held liable under either product liability laws or tort laws. Under product liability laws, developers shall be liable if the algorithm is treated as a defective product with flawed designs that inherently produces biased results. Under tort laws, developers shall be liable for negligence if they failed to exercise reasonable care such as ensuring the absence of bias in the training data or performing adequate testing and validation of the algorithm. Occasionally, Courts assign liability to both developers and deployers if both contributed to the harm.

In EU, it's emphasized to prioritize structured prevention and strict regulatory compliance with comprehensive frameworks such as the General Data Protection Regulation (GDPR) and the forthcoming Artificial Intelligence Act. Liability of bias is determined not only by harm caused but also by whether responsible parties complied with obligations regulated in laws. AI developers have major obligations to ensure that their systems are designed to avoid bias and meet legal requirements of safety, transparency, and fairness. If developers fail to comply, they shall be held liable. When algorithms create discriminatory impacts, deployers using AI systems shall be liable because they are obligated to avoid harmful automated decision-making without meaningful human oversight. Therefore, deployers have to actively take enough measures to make sure the achievement of fairness. Often, shared responsibility is imposed between AI developers and deployers to jointly be responsible for establishing safeguards of fairness. It's noted that non-compliance faces very high administrative fines.

5 Conclusion

With increasing uses of algorithms for hiring, the consequent algorithmic bias has inevitably frustrated the labor market. The belief that neutral algorithms replacing human decision-making can end human bias is ended by complexity of the AI world. As a fundamental right for human beings, the equal employment right shall be protected especially facing the challenge from monster of AI, which highlights the need to confront algorithmic bias at its root—whether it originates from the stage of data collection or the process of modelling. For such colorful sources of bias, a comprehensive approach including both technical and legal levels shall be seriously considered to try to mitigate them. It

is crucial to note that both technical and legal interventions need to remain dynamic and adaptable as the field of AI continues to evolve. The fast-paced nature of AI development reflects that measures once considered effective might eventually fail to address new challenges in algorithmic bias. Similarly, outdated legal frameworks may fail to regulate rapidly emerging AI applications. Therefore, the continuous technical and legal research is not only desirable but also necessary, which reflects the essence of human innovation and adaptability. After all, this is how the world runs, with the new instead of the old.

About the Author

Rui Zhong, Master Degree Candidate civil and commercial law, dispute resolution.

References

[1] Albaroudi E., Mansouri T., & Alameer A. (2024). A comprehensive review of AI techniques for addressing algorithmic bias in job hiring. AI, 5(1), 383–404.

[2] Chen Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. Humanities and Social Sciences Communications, 10(1).

[3] Gillis T. B., & Spiess J. (2019, August 2). *Big data and discrimination*. Papers.ssrn.com; The University of Chicago Law Review. https://papers.ssrn.com/sol3/papers.cfm? abstract_id=3204674.

[4] Hassija V., Chamola V., Mahapatra A., Singal A., Goel D., Huang K., Scardapane S., Spinelli I., Mahmud M., & Hussain A. (2023). Interpreting black-box models: A review on explainable artificial intelligence. Cognitive Computation, 16(1), 45–74.

[5] Hooker S. (2021). Moving beyond "algorithmic bias is a data problem." Patterns, 2(4), 100241.

[6] Houser K. (2019, February 28). Can AI solve the diversity problem in the tech industry? Mitigating noise and bias in employment decision-making. Papers.ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3344751.

[7] Kordzadeh N., & Ghasemaghaei M. (2021). Algorithmic bias: Review, synthesis, and future research directions. European Journal of Information Systems, 31(3), 1–22.

[8] Lytton C. (2024, February 16). AI hiring tools may be filtering out the best job applicants. BBC; BBC. https://www.bbc.com/worklife/article/20240214-ai-recruiting-hiring- software-bias-discrimination.

[9] Marieke A., & Sach A. (2024). Michael is better than mehmet: Exploring the perils of algorithmic biases and selective adherence to advice from automated decision support systems in hiring. Frontiers in Psychology, 15.

[10] Rafi M. (2024, October 14). When AI plays favourites: How algorithmic bias shapes the hiring process. The Conversation. https://theconversation.com/when-ai-plays- favourites-how-algorithmic-bias-shapes-the-hiring-process-239471.

[11] Rigotti C., & Fosch-Villaronga E. (2024). Fairness, AI & recruitment. Computer Law & Security Review, 53, 105966.

[12] Schellmann H. (2024, January 18). The algorithm: How AI can hijack your career and steal your future. Hurst Publishers.

[13] Society for Industrial, Organizational Psychology (US), and American Psychological Association. Division of Industrial-Organizational Psychology. (2018). Principles for the validation and use of personnel selection procedures. American Psychological Association, 11(S1), 1–97.

[14] The High-Level Expert Group on Artificial Intelligence. (2018, December 17). Ethics guidelines for trustworthy AI. FUTURIUM - European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html.

[15] Vasyl Rakivnenko. (2024, October 1). Toward responsible AI: Uncovering gender bias in leading embedding models. Forbes. https://www.forbes.com/councils/forbesbusinessdevelopmentcouncil/2024/10/01/toward-responsible-ai-uncovering-gender-bias-in-leading-embedding-models/.

[16] Wang X., Wu Y. C., Ji X., & Fu H. (2024). Algorithmic discrimination: Examining its types and regulatory measures with emphasis on US legal practices. Frontiers in Artificial Intelligence, 7.

[17] Yap J. Q., & Lim E. (2022). A legal framework for artificial intelligence fairness reporting. The Cambridge Law Journal, 81(3), 1–35.