Mitigating Bias in AI-Driven Hiring Systems for Manufacturing

CS4991 Capstone Report, 2025

Shevany Moharir
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
vnj3mf@virginia.edu

ABSTRACT

AI-driven hiring systems have significantly changed recruitment processes in the manufacturing industry over the past decade, but are prone to biases against groups such as displaced workers, older employees, and minorities. I propose a targeted framework to identify the causes of biases and ensure fair and inclusive hiring practices by incorporating fairness metrics to evaluate algorithmic bias, mitigation strategies such as re-weighting and adversarial debiasing to address disparities, and explainability tools for transparency in decision-making. Testing on simulated manufacturing recruitment datasets demonstrates improved fairness metrics without compromising model performance. Results show that tailored mitigation techniques can enhance workforce diversity and help connect displaced workers with reskilling opportunities. Future work includes scaling the framework to various real-world recruitment scenarios, addressing domain-specific challenges such as physical job requirements and regional disparities, and integrating compliance mechanisms for ethical and legal standards.

1. INTRODUCTION

In the manufacturing industries, artificial intelligence has rapidly emerged as a crucial hiring tool. By automating resume screening, candidate evaluation, and, more recently, job performance prediction, AI-driven recruitment systems seek to improve productivity. These algorithms may, however, make discriminatory hiring decisions against particular individuals since they rely on historical recruiting data that may contain unconscious biases. These biased hiring algorithms can further disadvantage older workers, displaced workers, racial and gender minorities, and other historically underrepresented groups in the manufacturing workforce, where automation is already significantly changing the business model.

Despite efforts to develop fairer AI models, many hiring algorithms continue to be affected by skewed training data, flawed feature selection, and black-box machine learning models. In manufacturing where labor shortages, skill mismatches, and displacement due to automation are pressing concerns, high stakes of recruitment decisions make it critical to address these biases. Having AI-based hiring tools that promote fair and inclusive practices is important for building a diverse and adaptable workforce that can navigate the evolving manufacturing industry. My proposal explores a targeted framework to detect and mitigate algorithmic bias in manufacturing recruitment, leveraging fairness metrics, mitigation techniques, and explainability tools to promote transparency and equitable hiring outcomes.

2. RELATED WORKS

The use of AI in hiring has been widely studied, highlighting both the potential benefits and risks. One of the most well-

known cases of algorithmic bias in hiring is Amazon's AI recruiting tool, which was found to systematically discriminate against female candidates due to historical hiring patterns (Goodman, 2018). The system learned to penalize resumes containing terms associated with women's colleges or activities, showing the implications of training AI models on biased data. Similar research discusses how algorithmic decision-making can unintentionally reinforce societal biases, particularly in hiring, where past hiring practices shape future predictions (Wang et al., 2024). These studies emphasize the need for fairness-aware machine learning techniques to ensure equitable outcomes.

Many methodologies have been proposed to mitigate AI bias in hiring. One approach is the concept of disparate impact removal, which aims to adjust feature distributions to minimize bias without significantly altering model performance (Wang et al., 2019). More recently, Kenna (2021) explored adversarial debiasing techniques, where a secondary model is trained to reduce discriminatory patterns in hiring decisions. These approaches provide a foundation for implementing fairness-aware algorithms that can be applied in the manufacturing sector to promote equitable hiring practices.

In addition to bias mitigation, there is also further research on explainability in AI hiring. Ribeiro et al. (2016) introduced Local Interpretable Model-agnostic Explanations (LIME), which provides insights into AI decision-making processes. A more recent study has explored SHAP (SHapley Additive exPlanations) as a tool for explaining model outputs in recruitment systems (Sogancioglu et al., 2023). These tools are critical in manufacturing recruitment, where transparency in hiring decisions can help address concerns of bias and promote trust in AI-driven hiring solutions. These studies lay the foundation for a comprehensive framework integrating bias detection, mitigation, and explainability tools tailored for manufacturing recruitment.

3. PROPOSAL DESIGN

My proposal outlines the development of a fairness-aware AI hiring framework for the manufacturing sector. The goal is to design a system that detects and mitigates biases in AI-driven hiring processes, encouraging transparency and accountability. The proposed solution will incorporate fairness metrics, bias mitigation techniques, and explainability tools to ensure equitable hiring practices in manufacturing.

3.1 Identifying Bias in AI-Driven Hiring

AI hiring software typically has biases inherent in historical employment data, which disproportionately harms some groups, such as displaced workers, older workers, and minorities. The first step in this proposal is to implement fairness metrics for measuring bias in AI hiring models. Disparate impact, equalized odds, and demographic parity will be used to measure imbalances in hiring recommendations. By evaluating these biases early in the pipeline, the system will deliver information on the potential discrimination ingrained in AI-driven decision-making.

One of the most important aspects of this phase is creating a diagnostic tool that critically examines hiring algorithms, pointing out where bias has been detected. This will include a fairness evaluation module to examine recruitment model output. The goal is to point out disparities by important demographic groups. This tool will serve as the foundation for bringing in remedial measures.

3.2 Bias Mitigation Strategies

To address identified biases, the proposal incorporates two strategies. The first is re-weighting techniques which includes adjusting training data weights to balance representation among demographic groups, ensuring fairer

outcomes. The second strategy is adversarial debiasing of training an auxiliary model to detect and penalize biased patterns in hiring decisions, discouraging reliance on sensitive attributes. These techniques aim to correct systemic imbalances without significantly compromising model performance. The effectiveness of each method will be tested using simulated recruitment datasets relevant to manufacturing roles.Planned deliverables are the implementation of bias mitigation strategies in hiring algorithms and a comparative analysis of fairness metrics before and after applying mitigation techniques.

3.3 Explainability and Transparency

One of the central issues with AI-driven hiring is decision-making opacity. This proposal incorporates explainability techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) in order to boost transparency (Salih et al., 2024). Such methods will make it possible for hiring managers to understand why a particular candidate was recommended and ensure that decisions can be audited for fairness. In addition, a prototype dashboard will be designed to visually present feature importance for hiring. This instrument will be capable of pinpointing key drivers of recruitment recommendations, enabling HR practitioners to estimate the impact of various attributes on recruitment outcomes, and ensuring accountability in AI-based hiring.

3.4 Plan and Future Considerations

This framework will initially be tested using synthetic recruitment datasets, followed by validation on real-world data where available. Future extensions will focus on scaling the framework to various manufacturing roles and industries and dressing domain-specific challenges, such as physical job requirements and regional hiring disparities. Further work could include integrating legal and ethical compliance measures to align with Equal Employment Opportunity Commission (EEOC) guidelines. By addressing these areas, the proposed system will ensure that AI-powered hiring in manufacturing promotes fairness, diversity, and inclusion while maintaining efficiency and reliability.

4. ANTICIPATED RESULTS

The proposed AI-powered recruitment system considers fairness and promotes diversity and inclusion in manufacturing by identifying and minimizing biases in recruitment systems. The majority of AI models inadvertently reinforce inequality toward disadvantaged displaced workers, aging workers, and underrepresented minorities. The inclusion of fairness considerations like disparate impact and equalized odds means the system will quantify bias, allowing HR professionals to correct for the imbalances. Techniques such as re-weighting and adversarial debiasing are expected to reduce disparities while maintaining predictive accuracy. Additionally, the incorporation of fairness constraints into model training will ensure compliance with regulatory codes and legislation, allowing companies to act proactively against discrimination in hiring. These interventions are designed to produce more equitable recruitment decisions without diminishing efficiency.

Along with technological innovations, the system also seeks to advance inclusive hiring practices that result in workforce diversity, innovation, and employee retention. A diverse workforce assists with productivity and problem-solving to the advantage of manufacturing companies. Explainability tools such as SHAP and LIME will provide transparency to AI-driven recommendations, allowing hiring managers to understand recommendations and trust automated

processes. Increased transparency can also improve connections with government agencies and unions, demonstrating a commitment to fair hiring practices. Ethical AI use can boost corporate image and public opinion, reassuring companies' dedication to diversity and inclusion.

5. CONCLUSION

The proposed fairness-aware AI hiring framework addresses critical biases embedded in hiring platforms within the manufacturing industry, making equitable hiring easier among displaced laborers, aged workers, and underrepresented minorities. Through the integration of fairness measures, mitigation strategies, and explainability techniques, the framework offers improved transparency and responsibility for AI-based hiring. Not only does the program promote labor force diversity but it also makes manufacturers stick to ethical and regulatory obligations. Ultimately, the system is designed to produce efficient and equitable hiring processes that contribute to a more inclusive workplace benefiting employers, job seekers and the economy as a whole.

6. FUTURE WORK

Future work will focus on refining the framework for real-world applications, addressing challenges such as physical job requirements, credential validation and regional labor conditions. Since manufacturing roles often demand technical skills and hands-on experience, adjustments may be necessary to prevent unintended bias while maintaining efficiency. Long-term studies will evaluate the impact of fairness interventions on hiring trends, workforce composition and employee retention. Additionally, expanding the framework beyond manufacturing could offer ethical AI recruitment solutions across industries, ensuring broader applicability.

Future improvements can also include adding mechanisms for compliance with evolving labor laws and ethical standards. Advances in AI interpretability can also enhance the ability of hiring managers to assess candidate recommendations, reinforcing trust and accountability in automated hiring. Through continual metric measurement and optimization of fairness, this solution has the potential to set a new standard for equitable and explainable AI-based hiring.

REFERENCES

Goodman, R. (2023). Why Amazon's automated hiring tool discriminated against women. ACLU. American Civil Liberties Union. https://www.aclu.org/news/womens-rights/why-amazons-automated-hiring-tool-discriminated-against

Kenna, D. (2021). Using adversarial debiasing to remove bias from Word embeddings. *ArXiv*. https://doi.org/10.48550/arXiv.2107.10251

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-Agnostic Interpretability of Machine Learning. *ArXiv*. https://doi.org/10.48550/arXiv.1606.05386

Salih, A., Raisi-Estabragh, Z., Galazzo, I., Radeva, P., Petersen, S., Lekadir, K. and Menegaz, G. (2024), A perspective on explainable artificial intelligence methods. SHAP and LIME. *Adv. Intell. Syst.*, 7: 2400304. https://doi.org/10.1002/aisy.202400304

Soğancıoğlu, G., Kaya, H., & Salah, A. A. (2023). Using explainability for bias mitigation: A case study for fair recruitment assessment. *Proceedings of the 25th International Conference on*

Multimodal Interaction: 631-639. https://doi.org/10.1145/3577190.3614170

Wang, H., Ustun, B., & Calmon, F. P. (2019). Repairing without retraining: Avoiding disparate impact with counterfactual distributions. *International Conference on Machine Learning*. https://doi.org/10.48550/arXiv.1901.10501

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*, 1320277. https://doi.org/10.3389/frai.2024.1320277