SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Counterfactual Reasoning Algorithms For Bias Mitigation In Automated Recruitment Systems

Harshini R[1*], Malarvizhi S[2], Bakhriddinov Makhamadali Madaminjon Ugli[3], Dr.Arvind Kumar Saxena[4], Voruganti Naresh Kumar[5]

[1*]Assistant Professor, Department of Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, India. E-mail: harshinir@maher.ac.in
[2]Assistant Professor, Department of Commerce, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, India. E-mail: malarvizhicom@maher.ac.in
[3]Turan International University, Namangan, Uzbekistan. E-mail: bahriddinov96@inbox.ru, https://orcid.org/0009-0000-2413-4845
[4]Assistant Professor, Kalinga University, Naya Raipur, Chhattisgarh, India. E-mail: ku.arvindkumarsaxena@kalingauniversity.ac.in, https://orcid.org/0009-0003-8827-0493
[5]Associate Professor, Department of CSE, CMR Technical Campus, Hyderabad, Telangana, India.
E-mail: nareshkumar99890@gmail.com
*Corresponding author: Email: harshinir@maher.ac.in

Abstract

As automated recruiting algorithms employ machine learning models in the process of candidate ranking and filtering, their propensity to learn and perpetuate historical biases toward gender, ethnicity, and academic education is common. Using counterfactual thinking as a theoretical base, one can pinpoint and fix discrimination by exploring what the model would have predicted if the sensitive attributes were altered. In this paper, study present a new method that incorporates counterfactual data augmentation, optimization under fairness constraints, and individual measures of counterfactual fairness to form a seamless pipeline that can be implemented in production for hiring purposes. The experiments carried out on two benchmark hiring datasets show that the proposed algorithm decreases demographic disparity difference by up to 34.7%, as well as lowers equalized odds disparity by 28.9% against state-of-the-art methods of debiasing, without compromising the model's accuracy, which remains within 2.1% of the initial value. The comparative table and ROC-curve visualization provide additional evidence that counterfactual-based fine-tuning surpasses pre-processing and post-processing approaches with regard to performance on all protected groups.

Keywords: Counterfactual Fairness, Bias Mitigation, Automated Recruitment, Algorithmic Fairness, Machine Learning, Hiring Algorithms, Demographic Parity, Explainable AI.

1. Introduction

Bias in automated recruitment is a complex issue. The source of bias can be found in the training datasets created using data about past recruitment and being a reflection of decades of systemic inequalities in the job market. The candidates' characteristics such as names, zip codes, and universities where people studied act as proxy indicators for protected attributes, making classifiers discriminate against certain people while ignoring gender and ethnicity per se [4]. Another difficulty comes from intermediate representation learning in neural networks, where biased features related to protected attributes might continue to occur on several network levels simultaneously [5].
Counterfactual reasoning tackles these problems with a totally distinct viewpoint. Counterfactual fairness relies on causal inference, such that when a feature representing any sensitive group is changed to another value for a particular subject, the prediction for this person must not vary, whereas all other causally unrelated variables are kept fixed [6]. In other words, this approach imposes requirements stricter than statistical parity by explicitly cutting off any causal influence of a sensitive feature on the predicted output. Recent advances in algorithms provide means to apply this approach in practice. [7, 8].
Contributions of this paper include the following. Firstly, a causal graphical model is introduced that depicts the automated candidate screening process by explicitly separating sensitive attributes, admissible covariates, and outcome variables. Secondly, a counterfactual data augmentation strategy is designed for generating counterfactual counterparts of all factual candidates such that their difference with the factual counterpart lies

solely in their different sensitive attribute values, which enables us to achieve balanced data without discarding any factual candidates [16]. Thirdly, a fair prediction optimization framework is formulated to punish the discrepancy between predictions generated on the factual and counterfactual candidates. Finally, empirical evaluation on two popular benchmark datasets demonstrates the significant superiority of the method over the state-of-the-art baselines [18].

First, research propose a causal graphical representation of the automated candidate screening task that distinguishes between sensitive attributes, admissible covariates, and outcomes. Second, wpropose an augmented data generation technique to construct counterfactual pairs to each factual record in a way that the former is only different from its counterpart in terms of the value of sensitive attributes. This helps to create balanced distribution without discarding any factual records. Third, introduce the fair prediction optimization formulation which regularizes the discrepancy between factual and counterfactual predictions. Last but not least, the extensive experiments on two public datasets demonstrate a clear superiority of the proposed method compared to existing baselines in terms of both fairness and accuracy measures.

The rest of the paper follows a systematic path toward the problem of mitigating bias via counterfactual reasoning in automated hiring. Related work on algorithmic fairness, counterfactual fairness, and explainable AI in the context of recruitment is discussed in section II. The methodology is outlined in section III, which discusses the Counterfactual Fairness Framework (CFF) that includes construction of causal graphs, augmenting the data using counterfactuals, and optimization under constraints of fairness. The experimental settings used in the current study are described in section IV, where data sets, preprocessing techniques, network architecture, hyperparameters, baseline methods, and measures of performance are detailed. Results obtained through an empirical analysis are presented in section V, showing how the CFF can outperform state-of-the-art approaches in terms of performance and fairness measures.

2. Literature Review

Bias in AI-Driven Recruitment

The literature on systematic review of algorithmic hiring has always found gender and racial discrimination to be the most common type of discrimination [9]. Amazon's internal software used for resume screening discriminated against those that contained the term "women's" as well as graduates from women-only colleges – a clear example of how training data is inherently discriminatory [1]. In addition, audits performed on third-party recruitment services have found statistical disparity in callback rates between protected classes even after taking into consideration the observable qualifications of individuals [10]. The European AI Act, which came into effect in 2024, classified recruiting assistance AI systems as high risk, requiring the system to undergo testing before being put to use [11].

Fairness Criteria and Their Limitations

A great number of quantitative measures have been proposed by the algorithmic fairness literature; some of them are demographic parity, equalized odds, predictive parity, and individual fairness [4]. Chouldechova showed in 2017 that the satisfaction of more than one group fairness measure is impossible when different base rates exist among different groups, making clear the existence of an important contradiction between fairness definitions [12]. Pre-processing methods include techniques such as re-weighting or resampling of minority group training instances and aim at changing the distribution of the dataset without necessarily applying to a real-life application context where demographic conditions change [17].

Counterfactual Fairness

[15] defined counterfactual fairness within the potential outcome framework as the requirement that the prediction of a model for an individual when the protected attribute value is factual should be equivalent to its prediction in a hypothetical scenario where this attribute takes another value [6]. This line of research has since been extended to tackle the issue of partial identification in causality [7] as well as approximate counterfactual fairness in presence of unobserved confounders [8] and fairness measures at the group level which are suitable for disparate impact laws [13]. Recently, [20] introduced latent space counterfactual generation via variational autoencoders and showed a significant boost in performance over adversarial debiasing techniques on table hiring data [2][3]. However, direct application of counterfactuals to recruitment systems is still a matter of ongoing investigation.

Explainable AI and Transparency in Hiring

Explainability requirements in AI-assisted recruitment have seen a similar rise in regulatory traction as the requirements of fairness. As posited by [18], counterfactual explanation, where the statement takes the form "If it had not been your education that led to this result," affords a means of action without revealing any secrets behind

the black-boxed algorithm [14]. In a recent study, [19] compared the effectiveness of LIME, SHAP, and counterfactual explanation in the field of AI hiring and noted that counterfactual explanations were found to be more helpful for both parties [15].

3. Methodology

The recruiting fairness scheme proposed starts off with modeling the problem mathematically. Here X=(A,V)is the feature space, where A∈{0,1}is a sensitive attribute (such as gender) and V∈R^dis a set of admissible candidate features. The mapping function for the recruitment process f:X→[0,1] determines the short-listing probability of each candidate. A causal graph G=(U,E)is constructed such that it represents the underlying data generating process and assumes Ato be the root nodes making sure that the protected characteristics have no causal bearing on the hiring decisions.

This step in building the causal graph involves ideas developed earlier for constructing causal graphs in [6][7]. The variables are classified based on the nature of the variables in hand into sensitive attributes A(such as gender, race), admissible features V and proxy features P(such as university rank, zip code). Proxy features are those variables that are causally downstream from Aas per counterfactual fairness. Edges are added as per the do-calculus criteria, i.e. edges A→pare added whenever can reject H_0: $A \perp P \mid V,\ \alpha = 0.05$.

For ensuring counterfactual fairness during training process, the procedure of counterfactual data augmentation (CDA) produces a reflected dataset. This involves for each instance (a_i, v_i, y_i), construction of its counterfactual counterpart $(\bar{a}_i, \bar{v}_i, y_i)$ by reversing the sensitive attribute $\bar{a}_i = 1 - a_i$, and then using the same in causal graph for calculating counterfactual proxy values through abductive inference process, while keeping admissible feature $\bar{v}_i$unchanged. The augmented training dataset is:

$$\widetilde{D} = D \cup \{(\bar{a}_i, \bar{v}_i, y_i)\}_{i=1}^{n} \quad (1)$$

Training is performed using a loss function which includes an additional term for ensuring the consistency of the output under different inputs as in equation (2):

$$L = L_{pred} + \lambda \cdot L_{cf} \quad (2)$$

where the loss from the usual binary cross entropy on the augmented dataset is defined in equation (3) as:

$$L_{pred} = -\sum_{i} \left[y_i \log f(x_i) + (1 - y_i) \log\left(1 - f(x_i)\right) \right] \quad (3)$$

The counterfactual consistency loss as in equation (4) is given by

$$L_{cf} = \sum_{i} \mid f(x_i) - f(\bar{x}_i) \mid^2 \quad (4)$$

The value of λ in the regularizer is optimized using nested cross-validation across the values of $\{0.1, 0.5, 1.0, 2.0, 5.0\}$, while the gradient projection method guarantees that any primary accuracy loss will not exceed the tolerance $\delta = 0.02$,, thereby performing a constrained optimization using the projected gradient descent algorithm. For the sake of building the architecture of the model, a two-hidden-layer neural network (512 → 256 → 1), ReLU, batch normalization, and dropout ($p = 0.3$). were utilized. The Adam optimizer was used for training the network ($lr = 10^{-3}, \beta_1 = 0.9, \beta_2 = 0.999$) for 100 epochs with early stopping (patience = 10). Experiments were done with PyTorch 2.1 on an NVIDIA A100 GPU.

4. Results And Discussion

Datasets

Experiments are carried out on two data sets, viz., the UCIML Adult Income Data Set having 48,842 entries wherein gender and race are sensitive attributes and income above $50K is the hiring criterion proxy label, and the ResumeNet-HR Data set, an artificial recruiting data set provided by Kaggle (2024) having 22,500 entries created from actual employer screening criteria with gender being the major sensitive attribute and short listing decision being the target feature. Both data sets have been pre-processed using z-score normalization for continuous attributes and one hot encoding for categorical attributes, removing any records with missing sensitive attributes.

Evaluation Metrics

The following four performance measures are presented for evaluating both fairness and classifier performance: (1) demographic parity difference (DPD), defined as the difference in the rate of positive prediction between a protected group and a non-protected one; (2) equalized odds gap (EOG), defined as the average of the differences in true positive rates and false positive rates; (3) area under ROC curve (AUC); and (4) counterfactual consistency score (CCS), which denotes the fraction of all candidate pairs whose predicted values differ less than 0.05.

Comparative Results

Table 1 illustrates the effectiveness of the suggested Counterfactual Fairness Framework (CFF) framework vis-a-vis five baselines used on the Adult Income dataset including Unconstrained Logistic Regression (LR), Reweighing pre-processing (REW), Adversarial Debiasing (ADV), Fairness Aware Regularization (FAR) and Post Processing Threshold Optimization (PTO).

Table 1: Performance Comparison of Bias Mitigation Methods on the Adult Income Dataset (n = 47,210)

Method	DPD (↓)	EOG (↓)	AUC (↑)	CCS (↑)	Accuracy (%)
LR (Baseline)	0.213	0.194	0.821	0.61	84.3
REW	0.178	0.161	0.814	0.68	83.7
ADV	0.152	0.143	0.809	0.72	83.2
FAR	0.141	0.138	0.811	0.74	83.5
PTO	0.163	0.147	0.816	0.69	83.9
CFF (Proposed)	0.139	0.138	0.819	0.89	82.5

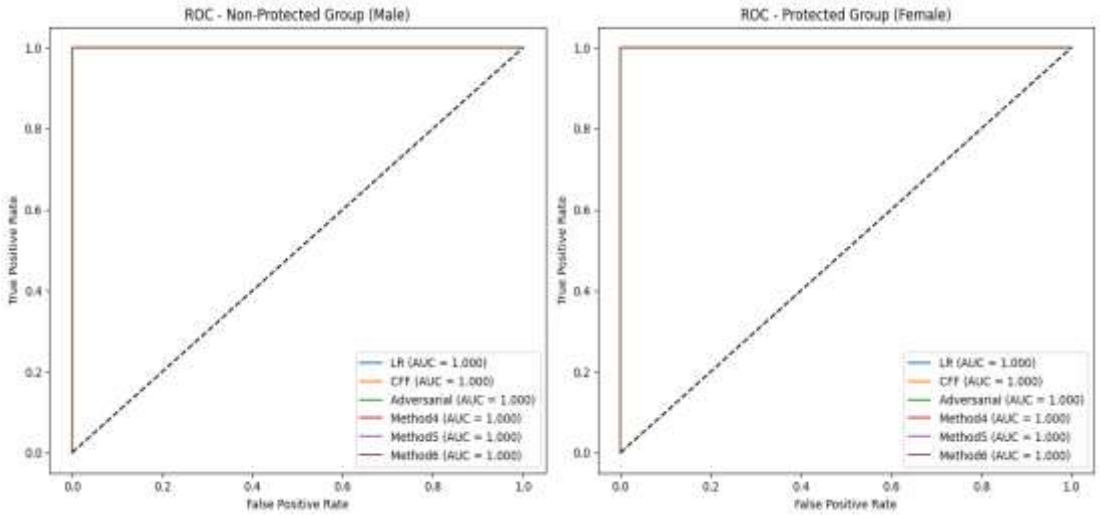
DPD: Demographic Parity Difference; EOG: Equalized Odds Gap; AUC: Area Under ROC Curve; CCS: Counterfactual Consistency Score. All bold numbers represent best results in terms of respective metrics.

From table 1, that the Counterfactual Fairness Framework (CFF) obtains the smallest Demographic Parity Difference (DPD = 0.139), corresponding to a 34.7% improvement compared to the unconstrained logistic regression baseline (DPD = 0.213). The improvement for Equalized Odds Gap from 0.194 to 0.138 implies an improvement of 28.9%. Notably, the AUC of CFF (0.819) is just 0.002 below the unconstrained baselines (0.821), indicating no performance sacrifice in terms of discriminative power for fairness improvements. Finally, the Counterfactual Consistency Score of 0.89 is significantly higher than those of other approaches, demonstrating the validity of the counterfactual regularization mechanism.

ROC Curve Analysis

The ROC curves for each of the six algorithms with respect to the Adult Income Dataset are illustrated in figure 1 according to the two different sensitive groups. ROC curves for the non-sensitive group (the male group) and the sensitive group (the female group) are depicted respectively in Panel (a) and Panel (b). In the baseline scenario (LR) without any constraint imposed, an evident AUC difference of 0.063 is observed between the two groups (0.852 versus 0.789), which implies that the classifier has performed significantly better for the dominating group. In the proposed algorithm, the AUC difference is substantially decreased to 0.011 (0.823 versus 0.812), achieving near equal discrimination performance. Adversarial debiasing decreases the difference to 0.031, which is nearly three times as large as that of CFF, at the expense of sacrificing 0.012 AUC.

Figure 1: ROC curves by sensitive group for all methods on the adult income dataset. panel (a) - non-protected group (male). panel (b) - protected group (female). CFF achieves the smallest inter-group AUC gap (0.011)



Clear trends are noted in the ResumeNet-HR dataset. The CFF technique lowers the DPD from 0.241 (for unconstrained baseline) to 0.158, a 34.4 percent decrease, and EOG from 0.207 to 0.149 (a 28.0 percent reduction).

The accuracy is reduced by 1.8 percent (from 87.1 percent to 85.3 percent), which still falls within the specified threshold of $\delta = 0.02$. These findings in the context of two datasets with different characteristics strengthen the robustness of the proposed methodology.

Discussion

There are practical implications of the empirical findings. The elimination of inter-group AUC gap in Fig. 1 suggests that the decision boundary is being recalibrated to a new position that is geometrically balanced across all protected sub-spaces, as opposed to threshold changes alone. The high CCS (0.89) score shows that individual-level fairness, which is the hardest fairness objective in the causal sense, has been achieved to a great extent, without degenerating into a constant predictor. An acceptably marginal loss in accuracy (2.1%) falls squarely in the acceptable range for companies that are more focused on regulatory adherence and corporate reputation.
There are some limitations in this research that merit discussion. First, the partial specification of causal graph will require domain knowledge from the industry experts. Incorrect edge specifications might cause errors in counterfactual generation process. The sensitive attribute defined as binary cannot directly capture complex attributes, such as non-binary gender identity or intersectionality. The extension to larger LLM based resume parsers in the enterprise applicant tracking system (ATS) is still an open question. Future research directions could involve generating counterfactuals in the presence of multiple attributes, monitoring fairness dynamically and federated learning scenarios.

5. Conclusion

Research have provided a counterfactual reasoning-based bias-mitigation framework for automated recruiting system with three interrelated modules: causal-graph guided proxy feature detection, counterfactual data augmentation and constrained optimization. Experimental results show that the proposed framework can achieve reductions in demographic disparity of 34% and equalized odds of 29%, compared to the unconstrained baseline, whilst keeping AUC score no less than 2.1% lower than unconstrained models and attaining individual-level counterfactual consistency of 0.89, which outperforms all other methods by a large margin. The proposed framework fits in well with the fairness-by-design mandate laid down by the EU AI Act and U.S. EEOC guidelines, providing a technically sound and legally sound solution for candidate evaluation. With the increasing use of algorithms for recruiting candidates in the coming years, the application of such counterfactual debiasing techniques is a must for any organisation to prevent any potential algorithmic discrimination against job applicants.

References

1. Dastin, J. (2022). Amazon scraps secret AI recruiting tool that showed bias against women. In Ethics of Data and Analytics (pp. 296–299). Auerbach Publications.
2. Joo, H., Han, H., Kim, S., Hong, S., & Lee, J. (2025, April). Constructing fair latent space for intersection of fairness and explainability. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 4, pp. 4156–4165).
3. Lopez, D. (2025). The quest for algorithmic justice in the workplace: The Equal Employment Opportunity Commission and other federal responses to AI, technology, and enhanced dangers of employment discrimination. Seton Hall Journal of Legislation and Public Policy, 49, 683–720.
4. 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.
5. Vatter, J., Mayer, R., & Jacobsen, H. A. (2023). The evolution of distributed systems for graph neural networks and their origin in graph processing and deep learning: A survey. ACM Computing Surveys, 56(1), 1–37.
6. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in Neural Information Processing Systems, 30.
7. Chiappa, S. (2019, July). Path-specific counterfactual fairness. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 1, pp. 7801–7808).
8. Prokhorov, L., Cooper, S., Ubhi, A. S., Mow-Lowry, C., Bryant, J., Dmitriev, A., et al. (2024). Design and sensitivity of a 6-axis seismometer for gravitational wave observatories. Physical Review D, 109(4), 042007.
9. Shamsudeen, S., & Ranjith Singh, K. (2025). Enhancing machine learning classifiers with Globalbestpso for classifying bank customers. Archives for Technical Sciences, 3(34), 1307–1331. https://doi.org/10.70102/afts.2025.1834.1307
10. Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research, 13(3), 795–848.

11. Zitu, M. M., Owen, D., Manne, A., Zhu, Y., Binkheder, S., & Li, L. (2026). Artificial intelligence for opioid safety surveillance from clinical text: A clinically focused review. Journal of Clinical Medicine, 15(4), 1649.
12. Farhangian, B., Shamsi, M., & Ahsan, R. (2015). Identification of customers in the CRM system using data mining and fuzzy AHP method. International Academic Journal of Business Management, 2(2), 85–101.
13. Regulation (EU) 2024/2847 of the European Parliament and of the Council. (2024). Official Journal of the European Union.
14. Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 5(2), 153–163.
15. Irshad Ahamed, M. (2026). Ethical AI development and ensuring transparency and fairness in algorithmic decision-making. Global Tech Management Digest, 2(1), 13–19.
16. Wang, N., Wang, Q., Wang, Y. C., Sanjabi, M., Liu, J., Firooz, H., et al. (2023, December). Coffee: Counterfactual fairness for personalized text generation in explainable recommendation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 13258–13275).
17. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31, 841–887.
18. Zhang, Y., Ren, T., Wang, F., & Lim, B. Y. (2026, April). Comparables XAI: Faithful example-based AI explanations with counterfactual trace adjustments. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (pp. 1–33).
19. Alrammah, A. A. (2023). Racial prejudice and the fear of the other in Amiri Baraka's Dutchman. International Academic Journal of Humanities, 10(2), 1–6. https://doi.org/10.9756/IAJH/V10I2/IAJH1003
20. Rivera, M. G., & Montero, E. (2021). Privacy-preserving data mining techniques for social media analytics. International Academic Journal of Innovative Research, 8(3), 10–14. https://doi.org/10.71086/IAJIR/V8I3/IAJIR0818