Selective Post-hoc Bias Mitigation via Counterfactual Sensitivity

Roti Islam Shithy*, Maria Stratigi, and Kostas Stefanidis

Data Science Research Centre, Tampere University, Finland

Abstract

Fairness concerns in recommender systems arise when predictions are sensitive to protected attributes, leading to unequal treatment of users and groups. While existing approaches often address either individual or group fairness through model retraining or constrained optimization, such methods can be difficult to deploy and audit in practice. In this work, a post-hoc counterfactual fairness framework has been proposed that jointly analyzes and mitigates individual- and group-level bias using a single, interpretable mechanism, without modifying or retraining the underlying model. This approach identifies unfair decisions through counterfactual sensitivity and applies a selective neutralization rule that removes the influence of sensitive attributes while preserving non-sensitive preference signals. Experiments on the MovieLens dataset using a logistic regression recommender demonstrate that the proposed method substantially reduces counterfactual sensitivity and demographic parity differences, while maintaining predictive accuracy. These results demonstrate that effective fairness auditing and mitigation can be achieved through simple, deployment-friendly post-hoc interventions.

Keywords

Counterfactual explanations, bias detection and mitigation, post-hoc fairness, explainable AI.

1. Introduction

Recommender systems increasingly shape user exposure and decision-making in online platforms. As a result, concerns about fairness have gained significant attention, particularly when recommendation outcomes depend on protected attributes such as gender, age, or occupation. Such dependencies may lead to unequal treatment of users or systematic disadvantages for certain demographic groups, raising ethical and societal concerns.

Fairness in recommender systems is challenging because unfair behavior can manifest at multiple levels. At the individual level, unfairness may arise when a user's recommendation outcome changes solely due to a modification of a protected attribute. At the group level, unfairness may appear as disparities in outcome distributions across demographic groups. While existing work has proposed numerous fairness-aware learning approaches, many focus on either individual or group fairness in isolation and often require model retraining, additional constraints, or complex optimization procedures.

Throughout this work, we operationalize unfairness as counterfactual sensitivity, defined as the extent to which a model's predictions change when protected attributes are altered while all other features remain fixed. Throughout the remainder of the paper, we use the term *bias* to refer to empirically measured unfairness, operationalized via counterfactual sensitivity at the individual level and its aggregation at the group level. This clarification allows us to consistently distinguish between the conceptual notion of unfairness and its measurable manifestations in experimental analysis.

Counterfactual sensitivity operationalizes a causal notion of individual fairness: a prediction is considered sensitive if it changes solely due to variation in a protected attribute while all other features remain fixed. Under this interpretation, fairness auditing becomes a test of whether protected attributes exert direct influence on model decisions, aligning with counterfactual definitions of individual fairness in prior work (Sandra Wachter, 2017); (Matt Kusner, 2017). This theoretical grounding clarifies why counterfactual sensitivity can serve as a principled signal for both bias detection and mitigation.

A large body of prior work addresses fairness during model training through regularization, adversarial learning, or constrained optimization (Alex Beutel, 2019); (Ninareh Mehrabi, 2021); (Joachims, 2018). Although effective, such in-

Published in the Proceedings of the Workshops of the EDBT/ICDT 2026 Joint Conference (March 24-27, 2026), Tampere, Finland
*Corresponding author.
ratiislamshithy@gmail.com (R. I. Shithy); maria.stratigi@tuni.fi (M. Stratigi); konstantinos.stefanidis@tuni.fi (K. Stefanidis)

0009-0003-1112-2935 (R. I. Shithy); 0000-0003-2482-4605 (M. Stratigi); 0000-0003-1317-8062 (K. Stefanidis)

processing approaches may be difficult to deploy in practice, especially in systems where models are already trained, proprietary, or costly to retrain. In contrast, post-hoc approaches aim to analyze and mitigate unfairness after deployment, offering greater flexibility and auditability (Sandra Wachter, 2017). However, existing post-hoc methods typically focus on either bias detection or bias mitigation and often emphasize a single notion of fairness (Ninareh Mehrabi, 2021); (Sandra Wachter, 2017).

In this paper, we propose a post-hoc counterfactual fairness framework that jointly analyzes and mitigates individual- and group-level bias using a single, interpretable mechanism. Our approach first identifies biased predictions through counterfactual sensitivity analysis and then applies a selective neutralization rule that removes the influence of protected attributes only when they affect the predicted outcome. Importantly, this process operates entirely at the prediction level and does not require retraining, constraint optimization, or modification of ranking objectives.

Unlike many fairness-aware recommender approaches that rely on model retraining or constraint-based optimization, this work adopts a post-hoc perspective that prioritizes interpretability and practical deployability. The focus is on lightweight analysis and intervention mechanisms that can be applied to existing systems without modifying their internal structure.

The contributions of this work are summarized as follows:

- We introduce a unified post-hoc framework that uses counterfactual sensitivity to analyze both individual- and group-level bias within the same mechanism.
- We propose a simple and selective neutralization strategy that mitigates biased predictions without retraining or modifying the underlying recommender model.
- We empirically demonstrate that the proposed approach reduces counterfactual sensitivity and group-level disparity on a real-world dataset, while preserving predictive accuracy.

2. Related Work

Fairness in recommender systems has been widely studied due to concerns that personalized recommendations may amplify data and societal biases. Early work showed that collaborative filtering and ranking-based methods can introduce popularity, exposure, and demographic biases, resulting in unequal treatment of users and items (T. Kamishima, 2011); (Burke, 2017); (Ekstrand, 2018). Such biases may disadvantage minority users and reinforce stereotypes, motivating fairness-aware recommendation research.

Fairness is commonly examined from two perspectives: group fairness and individual fairness. Group fairness focuses on reducing disparities across demographic groups and is often measured using demographic parity or related metrics (Asia J. Biega, 2018); (Ninareh Mehrabi, 2021). Individual fairness requires that similar users receive similar recommendations regardless of sensitive attributes (Sandra Wachter, 2017); (Matt Kusner, 2017). Prior studies note that optimizing one notion may negatively impact the other, highlighting the need for balanced approaches (Lucic, 2022).

Several methods address fairness during model training using regularization, adversarial learning, or constrained optimization (Joachims, 2018); (Alex Beutel, 2019). Although effective, these in-processing approaches often increase model complexity and reduce interpretability. Pre-processing techniques, such as data reweighting, attempt to correct biased distributions but may distort preference signals (Cowgill, 2020).

Post-hoc mitigation strategies provide a practical alternative by modifying predictions without retraining. These methods are particularly attractive in deployment settings and have been shown to improve fairness while preserving utility (Ninareh Mehrabi, 2021). Counterfactual explanations play a central role in post-hoc fairness auditing by assessing whether predictions change when sensitive attributes are altered while other features remain fixed (Sandra Wachter, 2017); (Dandl, 2020); (Poyiadzi, 2020); (Yash Goyal, 2019).

Recent work further improves counterfactual quality by optimizing feasibility and interpretability (Lucic, 2022). Fairness has also been explored beyond single-user recommendations. (Stefanidis, 2021) studied fairness in multi-round group recommendation, demonstrating how repeated interactions affect equity over time. (Stratigi, 2025) extended counterfactual reasoning to group recommendations using a model-agnostic framework, showing its effectiveness in heterogeneous settings.

Fairness in recommender systems has also been examined from broader and longer-term perspectives. (Eirini Ntoutsi, 2016) discussed early challenges related to bias and fairness in data mining and decision-making systems, highlighting how unfair outcomes may emerge and persist over time. Subsequent research has explored fairness beyond

single-shot recommendations, including temporal and long-term considerations. (Rodrigo Borges, 2020) investigated long-term fairness in recommender systems using variational autoencoders, illustrating how fairness objectives can be incorporated into representation learning over extended recommendation horizons. Similarly, the study (Rodrigo Borges, 2021) introduced variational autoencoder-based debiasing to mitigate popularity bias**,** showing that generative latent modelling can substantially reduce item over-exposure while maintaining ranking quality. More recent studies further examine fairness-related challenges in contemporary recommender and data-driven systems, including analyses of bias and fairness considerations in modern recommendation settings (Ali, 2025); (Hasan, 2024); (Dalla Vecchia, 2025).

Despite these advances, many existing approaches either (i) integrate fairness constraints during training, (ii) apply adversarial or optimization-based mitigation strategies, or (iii) focus exclusively on fairness auditing without a corresponding mitigation mechanism (Ekstrand, 2018); (Alex Beutel, 2019). In contrast, the proposed framework operates entirely at prediction time and uses counterfactual sensitivity as a single unifying mechanism for both bias detection and mitigation. Group-level effects are not enforced through explicit constraints but emerge from consistent removal of individual counterfactual sensitivity. This distinction positions the approach as a lightweight, deployment-friendly alternative to retraining-based or constraint-driven fairness methods.

3. Dataset and Preprocessing

3.1. Dataset

We conduct our experiments using the MovieLens 1M dataset, a widely used benchmark for recommender system research. The dataset contains 1,000,209 explicit ratings provided by 6,040 users on 3,900+ movies. In addition to user–item interactions, the dataset includes demographic attributes such as gender, age, and occupation, which are essential for fairness analysis across sensitive user groups.

The scale and demographic richness of MovieLens 1M make it suitable for evaluating bias detection and mitigation methods under realistic conditions, while providing sufficient data for stable group-level analysis.

3.2. Label Construction

The recommendation task is formulated as a binary classification problem. Explicit ratings are converted into binary labels following common practice in fairness-aware recommendation studies. Ratings greater than or equal to 4 are labeled as positive (liked), while ratings below 4 are labeled as negative (not liked). This binarization simplifies evaluation and enables direct comparison of positive outcome allocation across demographic groups.

3.3. Feature Representation

Each user–item interaction is represented by a feature vector combining user and item attributes. User features include gender, age, and occupation. Gender is encoded as a binary variable, while occupation is transformed using one-hot encoding to avoid ordinal assumptions. Age is retained as a numerical attribute to preserve demographic variation for fairness analysis. Item features consist of movie genres, represented using a multi-hot encoding scheme to capture multiple genre memberships. The final input vector is obtained by concatenating user and item features into a fixed-length representation suitable for linear classification models.

3.4. Data Splitting

The dataset is divided into training and test sets using an 80/20 stratified split. Stratification preserves the distribution of positive and negative labels as well as demographic attributes across splits. This setup ensures that fairness and accuracy evaluations before and after mitigation are not affected by sampling imbalance.

4. Methodology

This section describes the proposed post-hoc fairness framework for detecting and mitigating bias in recommender systems using counterfactual explanations. The framework is model-agnostic and operates entirely at the prediction level, preserving the original learning process.

4.1. Model Training

We employ logistic regression as the base recommender model due to its interpretability and efficiency. The task is formulated as binary classification, where the model predicts whether a user likes an item.

Let $X \in \mathbb{R}^{n \times d}$ denote the feature matrix and $y \in \{0,1\}^n$ the binary labels. The logistic regression model estimates:

$$\hat{y} = \sigma(X\beta), \quad (1)$$

Where $\sigma(\cdot)$ is the sigmoid function and β represents model parameters.

The dataset is split into training and test sets using an 80/20 stratified split. The trained model produces baseline predictions on the test set, which are later used for fairness analysis and mitigation.

4.2. Counterfactual Bias Detection

Bias is detected using counterfactual reasoning by evaluating prediction sensitivity to changes in sensitive attributes. For each test instance x_i, a counterfactual instance x_i' is generated by modifying only one sensitive attribute (e.g., gender) while keeping all other features unchanged. Counterfactual instances are generated by modifying exactly one protected attribute at a time while keeping all remaining features fixed. For binary attributes such as gender, the value is flipped (e.g., $0 \leftrightarrow 1$). For categorical attributes such as occupation, a single alternative valid category is assigned to construct the counterfactual instance. No optimization-based search or multiple counterfactual candidates are generated. The objective is to test local prediction sensitivity under minimal attribute intervention rather than to explore the full counterfactual space. This design ensures determinism and reproducibility of the sensitivity test.

4.2.1. Individual-level Bias

Individual-level bias is identified by comparing the model's predictions for x_i and its counterfactual x_i':

$$b_i = \begin{cases} 1, & \text{if } \hat{y}(x_i) \neq \hat{y}(x_i') \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

A prediction flip ($b_i = 1$) indicates that the model's decision for instance i depends on the sensitive attribute, suggesting individual-level unfairness.

4.2.2. Group-level Bias

Group-level bias is measured by aggregating individual flips within demographic groups. Let G denote a demographic group (e.g., female users). The group-level bias rate is defined as:

$$\rho_G = \frac{1}{|G|} \sum_{x_i \in G} \mathbb{I}[\hat{y}(x_i) \neq \hat{y}(x_i')], \quad (3)$$

Where $\mathbb{I}[\cdot]$ is the indicator function. Higher values of ρ_G indicate stronger sensitivity of model predictions to the sensitive attribute within group G. In reporting results, this bias rate is expressed as a percentage ($\rho_G \times 100$) to facilitate comparison across demographic groups. This metric captures systematic disparities that may not be visible through individual analysis alone.

4.3. Post-hoc Counterfactual Neutralization

To mitigate detected bias, we apply a post-hoc counterfactual neutralization strategy. For each test instance, the original prediction is compared with its counterfactual prediction.

The fairness-adjusted prediction $\hat{y}_{fair,i}$ is defined as:

$$\hat{y}_{fair,i} = \begin{cases} \hat{y}(x_i'), & \text{if } \hat{y}(x_i) \neq \hat{y}(x_i') \\ \hat{y}(x_i), & \text{otherwise} \end{cases} \quad (4)$$

The selective neutralization rule in Eq. (4) is intentionally designed as a minimal and deterministic post-hoc intervention. Rather than enforcing global fairness constraints or uniformly adjusting predictions, the rule intervenes only when a prediction is counterfactually sensitive to a protected attribute. Predictions that remain invariant under counterfactual changes are left unchanged. This design ensures that mitigation is targeted, interpretable, and avoids unnecessary distortion of non-sensitive preference signals, while preserving model utility and decision fidelity. Importantly, the approach can be applied without retraining or modifying the underlying recommender system.

No group-specific optimization is performed. Group-level bias reductions emerge solely from the application of the same selective neutralization rule to individual predictions, followed by aggregation across demographic groups. As a result, bias detection and mitigation are unified through a single counterfactual sensitivity signal, rather than treated as separate stages.

A natural alternative to post-hoc mitigation is the removal of sensitive features or the unconditional use of counterfactual predictions. However, feature removal does not guarantee fairness, as protected attributes may remain implicitly encoded through correlated features, and it prevents auditing whether individual decisions are sensitive to protected attributes. Similarly, always switching to counterfactual predictions or applying demographic post-processing would uniformly alter outcomes, including those that are already unbiased. In contrast, selective neutralization intervenes only when counterfactual inconsistency is detected, enabling targeted and auditable mitigation.

In addition to counterfactual sensitivity, we report standard group-level fairness indicators, namely

positive outcome rates (POR) and demographic parity difference (DPD), computed before and after post-hoc mitigation.

4.4. Evaluation Metrics

The proposed framework is evaluated using both fairness and utility metrics. Fairness is primarily assessed through counterfactual bias, measured as the rate of prediction changes when protected attributes are counterfactually modified while all other features remain fixed. This analysis is conducted at both the individual level and the group level across demographic attributes including gender, age, and occupation. To complement counterfactual bias, we also report standard group-level fairness indicators, positive outcome rates (POR) and demographic parity difference (DPD), before and after post-hoc mitigation. Model utility is evaluated using classification accuracy on the test set. All metrics are computed before and after selective neutralization to examine the resulting fairness–utility trade-off.

5. Experimental Results

This section presents the empirical evaluation of the proposed post-hoc counterfactual fairness framework on the MovieLens 1M dataset. We analyze bias at both the individual and group levels before and after selective neutralization, followed by an analytical summary of absolute and relative bias reductions. We further examine the impact of mitigation on model utility using prediction accuracy, and report standard group-level fairness indicators, including positive outcome rates (POR) and demographic parity difference (DPD), to provide a comprehensive assessment of fairness–utility trade-offs.

5.1. Individual-level Bias

Individual-level bias is measured as the proportion of test instances whose predicted outcome changes when gender is counterfactually flipped.

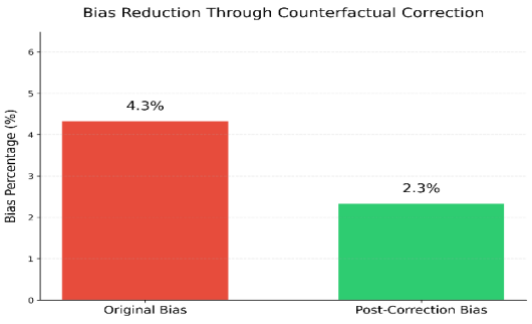

Figure 1: Individual-level bias before and after post-hoc counterfactual neutralization. Bias is measured as the proportion of users whose predictions change under a counterfactual gender flip. A bar chart with two bars. The first bar shows individual-level bias before mitigation at 4.3%. The second bar shows individual-level bias after mitigation at 2.3%, indicating a reduction in prediction changes due to gender.

5.2. Group-level Bias

Group-level bias was analyzed across age, gender, and occupation by aggregating counterfactual prediction flips within each demographic group.

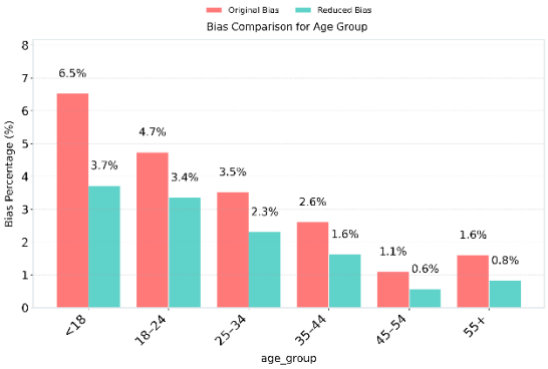

Figure 2: Group-level bias by age group before and after counterfactual neutralization. A grouped bar chart showing bias percentages for multiple age groups. Each age group has two bars representing bias before and after mitigation. Younger age group (<18) show higher baseline bias (6.5%) with consistent reductions (3.7%) after mitigation, and all other groups also show lower bias after mitigation.

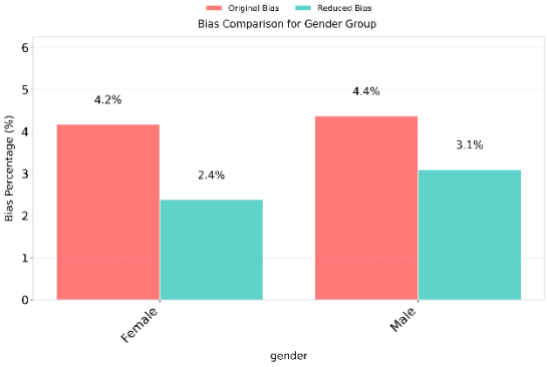

Figure 3: Group-level bias comparison between male and female users before and after counterfactual neutralization. A grouped bar chart with two categories: male and female users. Each category contains two bars representing bias before and after mitigation. Bias values are similar before mitigation (for male and female users 4.4% and 4.2%, respectively) and decrease for both groups afterward (for male and female users 3.1% and 2.4%, respectively).

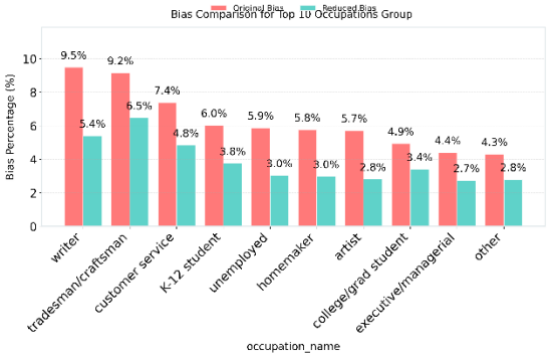

Figure 4: Group-level bias across occupational categories before and after post-hoc counterfactual neutralization. A grouped bar chart showing bias rates for several occupation categories. Each occupation has two bars indicating bias before and after mitigation. Occupations with higher initial bias show noticeable reductions after mitigation (for writers and tradesmen/craftsmen, bias dropped from 9.5% to 5.4% and 9.2% to 6.5% respectively), and bias is substantially reduced across all other occupational groups as well.

While the figures above illustrate reductions in individual and group-level bias after post-hoc neutralization, they do not directly convey the magnitude of these changes. To provide an analytical summary, we next report counterfactual bias levels before and after mitigation, together with absolute and relative reductions, aggregated across individual predictions and demographic attributes.

Table 1
Summary of counterfactual bias before and after post-hoc mitigation.

Attribute	Bias Before (%)	Bias After (%)	Absolute Reduction (%)	Relative Reduction (%)
Individual	4.32	2.33	1.99	46.06
Gender Group	4.27	2.74	1.53	35.94
Age Group	4.20	2.61	1.59	37.82
Occupation Group	4.23	2.47	1.75	41.52

These results show consistent bias reduction across individual predictions and group levels. Relative reductions range from 35% to 46%, with larger absolute reductions observed for attributes exhibiting higher baseline bias. This pattern indicates that the selective neutralization mechanism primarily affects predictions with strong counterfactual sensitivity, resulting in targeted mitigation rather than uniform adjustment across all users.

5.3. Accuracy

To assess the impact of post-hoc mitigation on model utility, prediction accuracy before and after selective neutralization has been reported.

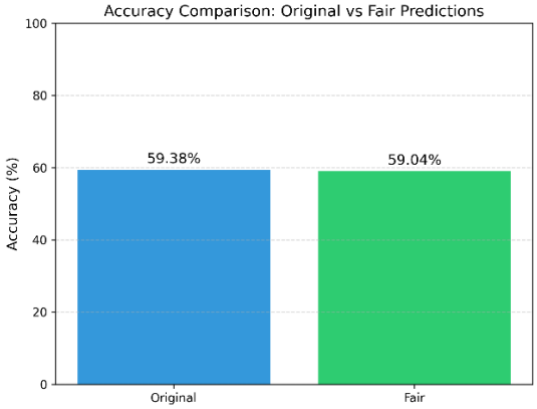

Figure 5: Accuracy comparison between original and fairness-adjusted predictions on the MovieLens 1M dataset. The figure shows a bar chart comparing two accuracy values. The first bar, labeled "Original," has an accuracy of 59.38%. The second bar, labeled "Fair," has an accuracy of 59.04%. The bars are similar in height, indicating minimal utility loss (0.34%) after bias mitigation.

5.4. Fairness Indicators (POR, DPD)

In addition, we report standard group-level fairness indicators to further assess mitigation impact. Specifically, we examine demographic parity difference (DPD) and positive outcome rates (POR) before and after mitigation, which provide complementary perspectives on changes in outcome distributions across demographic groups. These findings are summarized in Table 2 and Table 3.

Table 2
Group-level Demographic Parity Difference (DPD) before and after Post-hoc mitigation.

Attribute	DPD Before	DPD After	Absolute Reduction
Gender	0.0744	0.0110	0.0634
Age	0.2058	0.1850	0.0208
Occupation	0.5092	0.5015	0.0077

The results indicate DPD decreases across all attributes after selective neutralization, with the largest reduction observed for gender and more moderate decreases for age and occupation. This indicates reduced group-level disparity without explicit group-level constraints.

Table 3
Positive Outcome Rates (POR) before and after post-hoc mitigation across demographic groups.

Attribute	Group	POR Before (%)	POR After (%)
Gender	Male	80.83	85.20
	Female	88.27	84.10
Age	<18	73.79	76.94
	18-24	81.43	84.35
	25-34	85.64	86.95
	35-44	90.15	91.58
	45-54	92.38	92.94
	55+	94.37	95.44
Occupation	homemaker	88.46	82.69
	executive /managerial	86.03	89.28
	artist	82.45	83.83
	college / grad student	82.21	85.46
	K-12 student	79.74	82.33
	other	76.58	78.51
	customer service	70.43	76.96
	Tradesman/craftsman	69.72	78.87
	writer	66.27	71.36
	unemployed	45.75	49.02

6. Discussion

This work demonstrates that counterfactual sensitivity can serve as a unified mechanism for both fairness auditing and mitigation at prediction time. By applying a deterministic selective neutralization rule, biased predictions are adjusted without retraining or optimization, while invariant predictions remain unchanged.

The framework is model-agnostic and requires only prediction-level access. While logistic regression enhances transparency of prediction behavior, the counterfactual sensitivity test and selective neutralization rule remain applicable to more complex architectures, where interpretability refers primarily to the transparency of the mitigation mechanism rather than full model introspection.

We intentionally focus on an interpretable model (logistic regression) evaluated on a widely used benchmark dataset (MovieLens 1M) to isolate the effect of post-hoc counterfactual mitigation from confounding factors introduced by model complexity. This design choice allows observed fairness improvements to be attributed directly to the proposed intervention and aligns with practical deployment scenarios where retraining or architectural modification is undesirable.

Results show consistent reductions in individual and group bias with minimal accuracy impact, indicating a favorable fairness–utility trade-off. The proportion of modified predictions corresponds to the counterfactual sensitivity rate; thus, only instances exhibiting prediction flips under protected attribute variation are adjusted, while the majority of predictions remain unchanged. Complementary group-level indicators further show POR shift across demographic groups, while DPD is reduced after mitigation. These patterns suggest that selective neutralization performs targeted corrections for groups with higher baseline bias rather than uniformly altering outcomes.

This study has several limitations. The evaluation is restricted to a binary relevance setting, a single dataset, and one interpretable model, and counterfactual analysis is applied to one sensitive attribute at a time. Additionally, modifying a protected attribute while keeping all other features fixed may generate user profiles that are less plausible in practice. In this work, such counterfactuals are used strictly as diagnostic probes for sensitivity analysis rather than as realistic simulations of user transitions. Moreover, group-level metrics do not capture longer-term exposure or feedback dynamics. Furthermore, the proposed approach is not empirically compared against state-of-the-art in-processing or post-hoc mitigation methods. The objective of this study is to demonstrate the feasibility and behavior of a unified counterfactual sensitivity mechanism rather than to establish empirical superiority over optimization-based fairness techniques. Future work will extend the framework to ranking-based recommendations, more complex models, intersectional fairness settings, and longitudinal evaluations of post-hoc mitigation effects.

References

[1] Alex Beutel, J. C. (2019). Fairness in Recommendation Ranking through Pairwise Comparisons. New York, NY, USA: Association for Computing Machinery. doi:10.1145/3292500.3330745

[2] Ali, M. S. (2025). Fairness in Group Recommender Systems Using Variational Autoencoders. *IDEAS 2024. 15511* . Springer, Cham. doi:https://doi.org/10.1007/978-3-031-83472-1_20

[3] Asia J. Biega, K. P. (2018). Equity of Attention: Amortizing Individual Fairness in Rankings. New York, NY, USA: {Association for Computing Machinery. doi:10.1145/3209978.3210063

[4] Burke, R. (2017). Multisided Fairness for Recommendation. doi:10.48550/arXiv.1707.00093

[5] Cowgill, B. &. (2020). Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics. ResearchGate. doi:10.48550/arXiv.2012.02394

[6] Dalla Vecchia, A. &. (2025). Bias Evaluation in Contextual Machine Learning. doi:10.1007/978-3-032-05727-3_45.

[7] Dandl, S. M. (2020, August. 31). Multi-Objective Counterfactual Explanations. *Springer Nature Link*. doi:https://doi.org/10.1007/978-3-030-58112-1_31

[8] Eirini Ntoutsi, K. S. (2016). Recommendations beyond the ratings matrix. . *In Proceedings of the Workshop on Data-Driven Innovation on the Web (DDI '16)* (pp. 1-5). New York, NY, USA: Association for Computing Machinery . doi:10.1145/2911187.2914580

[9] Ekstrand, M. &. (2018). All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. *1st Conference on Fairness, Accountability and Transparency*. Retrieved from https://proceedings.mlr.press/v81/ekstrand18b.html

[10] Hasan, M. M. (2024). SQUIRREL 2.0: Fairness & Explanations for Sequential Group Recommendations. In O. L. Proceedings of the 26th International Workshop on Design (Ed.), *CEUR Workshop Proceedings. 3653*, pp. 63-67. Paestum: CEUR-WS.

[11] Joachims, A. S. (2018). Fairness of Exposure in Rankings. New York, NY, USA: Association for Computing Machinery. doi:10.1145/3219819.3220088

[12] Lucic, A. t. (2022). CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022).* PMLR (Proceedings of Machine Learning Research). doi:https://doi.org/10.48550/arXiv.2102.03322

[13] Matt Kusner, J. L. (2017). Counterfactual fairness. *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).* (pp. 4069-4079). ACM Digital Libraray. Retrieved from https://dl.acm.org/doi/10.5555/3294996.3295162

[14] Ninareh Mehrabi, F. M. (2021, July 13). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys (CSUR), 54*(6), pp. 1-35. doi:https://doi.org/10.1145/3457607

[15] Poyiadzi, R. &.-R. (2020). FACE: Feasible and Actionable Counterfactual Explanations. *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society* (pp. 344-350). ResearchGate. doi:10.1145/3375627.3375850

[16] Rodrigo Borges, K. Stefanidis (2020). Enhancing Long Term Fairness in Recommendations with Variational Autoencoders. *In Proceedings of the 11th International Conference on Management of Digital EcoSystems (MEDES '19)* (pp. 95-102). New York, NY, USA: Association for Computing Machinery. doi:https://doi.org/10.1145/3297662.336579

[17] Rodrigo Borges, K. Stefanidis (2021). On mitigating popularity bias in recommendations via variational autoencoders. In P. o. '21). (Ed.). (pp. 1383–1389). New York, NY, USA: Association for Computing Machinery. doi:https://doi.org/10.1145/3412841.344212

[18] Sandra Wachter, B. M. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology, 2018*. doi:https://doi.org/10.48550/arXiv.1711.0039

[19] Stefanidis, H. a. (2021). Multi-Round Recommendations for Stable Groups. *2021 IEEE International Conference on Progress in Informatics and Computing (PIC)* (pp. 232-240). Shanghai, China : IEEE Xplore. doi:10.1109/PIC53636.2021.9687062

[20] Stratigi, M. &. (2025). Counterfactual Explanations for Group Recommendations. *27th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP 2025).* ResearchGate.

[21] T. Kamishima, S. A. (2011). Fairness-aware Learning through Regularization Approach. *2011 IEEE 11th International Conference on Data Mining Workshops* (pp. 643-650). IEEE. doi:10.1109/ICDMW.2011.83.

[22] Yash Goyal, Z. W. (2019). Counterfactual Visual Explanations. *Proceedings of the 36th International Conference on Machine Learning* (pp. 2376–2384). PMLR. Retrieved from https://proceedings.mlr.press/v97/goyal19a.html

[23] Yifan Wang, W. M. (2023, July). A Survey on the Fairness of Recommender Systems. *ACM Transactions on Information Systems, 41*(3), pp. 1-43. Retrieved from https://doi.org/10.1145/3547333