Bias vs Bias Dawn of Justice: A Fair Fight in Recommendation Systems

Tahsin Alamgir Kheya(✉), Mohamed Reda Bouadjenek, and Sunil Aryal

Deakin University, Geelong, VIC, Australia
{t.kheya,reda.bouadjenek,sunil.aryal}@deakin.edu.au

Abstract. Recommendation systems play a crucial role in our daily lives by impacting user experience across various domains, including e-commerce, job advertisements, entertainment, etc. Given the vital role of such systems in our lives, practitioners must ensure they do not produce unfair and imbalanced recommendations. Previous work addressing bias in recommendations overlooked bias in certain item categories, potentially leaving some biases unaddressed. Additionally, most previous work on fair re-ranking focused on binary-sensitive attributes. In this paper, we address these issues by proposing a fairness-aware re-ranking approach that helps mitigate bias in different categories of items. This re-ranking approach leverages existing biases to correct disparities in recommendations across various demographic groups. We show how our approach can mitigate bias on multiple sensitive attributes, including gender, age, and occupation. We experimented on three real-world datasets to evaluate the effectiveness of our re-ranking scheme in mitigating bias in recommendations. Our results show how this approach helps mitigate social bias with little to no degradation in performance.

Keywords: Recommendation System · Fair re-ranking · Bias in Recommendations

1 Introduction

Recently, Recommendation Systems (RSs) have become an integral part of our lives by providing personalized suggestions to us. They play an important role in shaping our digital experience, contributing to our decisions for online purchases, movie recommendations, music playlists, news feeds, and more. RS spares us the trouble of sifting through vast amounts of data by curating customized and diverse content. Given their profound impact on our daily lives, it is essential to ensure they provide fair recommendations and do not perpetuate harmful biases. For instance, [58] highlights how top-ranked results for job roles favor one gender over the other and systematically disadvantage minority groups. While significant progress has been made in addressing fairness in recommendations [3,44,51], it is still an ongoing topic of research with new studies emerging continuously.

AI models are vulnerable to picking up biases that exist in the dataset used to train them [4,9,13,27,36]. For instance, [28] discusses how there is significant bias in movie recommendations for male and female users, across genres

R. P. Ribeiro et al. (Eds.): ECML PKDD 2025, LNAI 16013, pp. 248–264, 2026.
https://doi.org/10.1007/978-3-032-05962-8_15

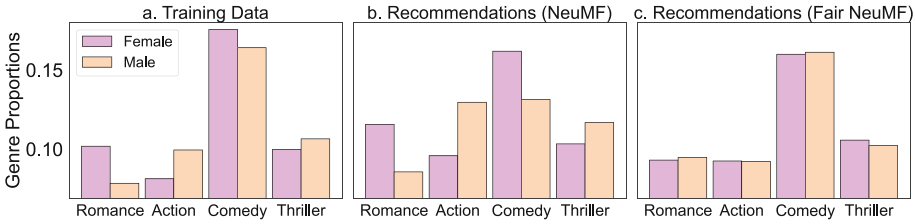

Fig. 1. Comparison of proportions of different movie genres for users of two genders in the training data, plain recommendations, and fairness-aware recommendations. There are disparities in the number of movies recommended to each gender for the four genres. This graph is based on the NeuMF [20] model for the ML100K dataset [19].

like romance and action. To mitigate such biases, researchers have used various fairness constraints to design re-ranking algorithms [15–17,58]. While there has been abundant work in the field of fair re-ranking in recommendations, existing approaches are deficient in two key ways: (i) they primarily focus on single or binary sensitive attributes and, (ii) they do not include the item categories when designing their approaches, which plays a vital role in users' end experience.

To illustrate the importance of mitigating bias on a granular scale by considering categories, we refer to Fig. 1. In this figure, we present how the proportions of movies from different genres vary for male and female users using the ML100K dataset across three stages: the training set, the top 20 plain recommendations provided by the NeuMF model, and the top 20 recommendations after fair re-ranking. From this figure, we can derive and explain several key insights: (i) The dataset used to train the model is not fair or neutral as shown in plot a. Certain genders tend to exhibit inherent biases towards specific categories, which societal stereotypes can influence. (ii) During the training process, the model can learn these biases and amplify them, intensifying their impact on the final output. We choose to demonstrate this using one popular recommendation model called NeuMF (plot b). The amplification of such biases by recommendation models is very common, as noted by [34,37]. While this helps visualize the discrepancy in movie recommendations for different genders, similar biases can exist in other domains, such as news recommendations, which can have far more profound consequences if not addressed. A similar concern is highlighted by [57], where YouTube recommender systems are found to facilitate pathways to extremist or radicalizing content. These kinds of content, when exposed to users, especially young generations, may have negative impacts on their well-being. (iii) The biases present in plots a and b, have similar trends of being oppositely skewed. As such, male users are more biased towards action and thrillers, while less towards romance and comedies. And female users are more biased towards romance and comedies while less towards action and thrillers. We take advantage of these opposing biases and use them as a corrective mechanism against social bias. After applying our fair category-aware re-ranking approach the discrep-

ancy in the categories of movies recommended to the different groups of users decreases significantly as seen in plot c.

Ensuring the categories of items are considered when designing a re-ranking scheme that also caters to multi-valued sensitive attributes is thus vital. This kind of refined approach will ensure a balanced distribution of recommended categories of items for different groups of users. In this paper, we introduce a fairness-aware re-ranking scheme that allows us to produce fair recommendations by considering users' social attributes, accommodating both binary and multi-valued attributes. This strategy builds on the concept of counterfactual fairness by leveraging the bias in the training set to tackle/counteract social bias in a category-aware setting. Essentially, we use the category preferences of users with different sensitive attributes to adjust the recommendations, leveraging the opposing biases to promote fairness. We evaluate the effectiveness of this scheme on different recommendation algorithms (including traditional and deep approaches), experimenting on three real-world datasets.

2 Related Work

2.1 Consumer-Side Fairness in Recommendation Systems

Fairness in recommendation systems is a multi-sided concept, which is categorized into (i) Provider-Fairness: fairness for providers or sellers in terms of exposure; (ii) Consumer-Fairness: which focuses on the fairness of items being recommended to users from different protected classes and (iii) CP-Fairness which considers both [6]. Our work focuses on C-Fairness, with our goal of similar recommendations regardless of the user's sensitive attributes. It has been observed in prior research how recommendation systems are prone to bias influenced by demographic factors like gender [8,11,12,39,42], age [12,42], occupation [32,54], race [49] and more. To promote fairness for consumers, researchers have proposed a variety of mitigation strategies that span across the pre-processing, in-processing, and post-processing stages of the ML pipeline. For instance, [45] shows how small additions of augmented data can substantially improve both individual and group fairness in recommender systems. The authors in [52] propose a multi-task adversarial learning scheme that satisfies three different fairness criteria, including group, individual, and counterfactual. Optimizing a fairness-aware regularization term along with the main recommendation loss is also a popular approach to mitigating bias in recommendations [2,5,28,53,56].

2.2 Fair Re-ranking

Re-ranking is a popular post-processing strategy to mitigate bias in recommendations. This method focuses on rearranging the items recommended to users for the top-k list by considering both recommendation quality and a fairness constraint. For instance, [15] introduces a re-ranking approach that incorporates a fairness constraint to mitigate unfairness in explainable recommenders that

use knowledge graphs. The authors in [31] introduce a fairness-constrained re-ranking method to ensure the utility disparity between different groups of users are below a certain threshold ϵ, while the optimization maximizes preference scores of items selected. Singh and Joachims [47] integrate common fairness concepts like demographic parity, disparate impact, and disparate treatment into their optimal ranking algorithm. The main idea behind the most relevant work is optimizing fairness and utility jointly by using a hyper-parameter to control the trade-off [51]. Although the current research community has not explored fair re-ranking for multi-valued attributes as much, there are a few works we wanted to highlight [22,38,50,55]. Unlike these, we are enforcing C-Fairness for multi-valued attributes in a category aware-setting. Most existing fair ranking schemes for recommendations focus on binary sensitive attributes and apply fairness definitions using the intuition of Equalized odds and Demographic parity to design their fairness constraint [2,31,40,47]. For demographic parity each sensitive group (like male and female) should receive the same proportions of positive predictions [18]. On the other hand, the concept of Equalized odds holds if the system has similar true positive rates and false positive rates across two different demographic groups [18]. In the current literature, these works would try to minimize the disparity between two groups of users or items based on popularity or user-sensitive attributes with binary values (like binary gender: [male, female], or age: [old, young]). While this approach is valuable in some way, it tends to oversimplify the complexity of user identities that are multi-dimensional.

Additionally, most re-ranking schemes don't consider the proportion of categories in the items recommended. Such schemes can fail to mitigate bias and disparities that exist across different types of items (like genres for movies). There are however some works that do consider different classes when designing ranking schemes [16,17,24,48]. For instance, [16] introduces algorithms to re-rank job candidates to achieve a desired distribution in the top results in regards to users' sensitive attributes like gender and age. The works by [16,17,24] aim for provider-side fairness. Our work is very close to that of [48], which re-ranks movies to ensure the recommendations align well with the historical interaction of the users by using genre distributions of previously played movies. Although this work is generating fair recommendations by ensuring users get recommendations following the proportions of genres that they previously watched, our work focuses on fairness in terms of users' sensitive attributes.

3 Proposed Re-ranking Scheme

3.1 Notation

We present all metrics for fairness assessment using the mathematical notation presented in Table 1.

We start with the two main distributions, both of which consider categories of the items.

Table 1. Notation Table

Notation	Description
$\mathcal{U}$ and $\mathcal{V}$	The set of users and items, respectively.
v_j	A single item, where j indexes the items.
c	An item category, such as Action, Sci-Fi, Romance, etc.
$\mathcal{C}$	The list of unique categories associated with all items.
C	A category matrix where $C_{v,c} = 1$ if item v belongs to category c, and 0 otherwise.
C_v	The list of categories associated with item v.
$\mathcal{V}_u$	The set of items the user u has interacted with in the past.
$t_{v,u}$	The timestamp of the interaction with item v by user u.
s_u	Represents the value of a sensitive attribute (male, female, engineer, etc.) for user u.
S	Represents a sensitive attribute like age, gender, occupation, etc.
$\text{score}_{u,v}$	The predicted score of item v by user u.

Definition 1 (Counterfactual Category Proportion (CCP)). *Let $o(c|s_u)$ return the average proportion of category c for all users who have a sensitive attribute that is not s_u, where $\mathcal{U}_{\neg s_u} = \{w \in \mathcal{U} \mid s_u \neq s_w\}$.*

$$o(c|s_u) = \frac{1}{|\mathcal{U}_{\neg s_u}|} \sum_{u \in \mathcal{U}_{\neg s_u}} m(c|u) \tag{1}$$

where

$$m(c|u) = \frac{\sum_{v \in \mathcal{V}_u} \frac{C_{v,c}}{|C_v|} \cdot t_{v,u}}{\sum_{v \in \mathcal{V}_u} t_{v,u}}$$

We use the timestamps of the interactions to apply more weight to interactions that took place recently. We follow [40], where they also employed the training set for their re-ranking algorithm.

Definition 2 (Recommended Category Proportion (RCP)). *Let $r(c|u)$ be the category proportion for user u relative to the items they are being recommended (represented by I) for category c.*

$$r(c|u, I) = \frac{\sum_{j=1}^{|I|} \frac{C_{v_j,c}}{|C_{v_j}|} \cdot \frac{1}{j^\gamma}}{\sum_{j=1}^{|I|} \frac{1}{j^\gamma}} \tag{2}$$

where we use $\gamma \in [0, 1]$ to help us weigh the item category contribution according to the rank (j) of the item in the recommended list I

3.2 Counterfactual Fairness

To design our category-aware fair re-ranking scheme, we use the concept of counterfactual fairness [30] that is formally defined as:

$$P\big(\hat{Y}_{A \leftarrow a}(U) = y \mid X = x, A = a\big) = P\big(\hat{Y}_{A \leftarrow a'}(U) = y \mid X = x, A = a\big)$$

Here, counterfactual fairness is achieved if the predicted outcome $\hat{Y}$ for an individual u (with latent variable U and non-sensitive features X) is the same when

intervening to externally set the user's sensitive attribute from a to a'. This concept avoids discrimination by making sure that sensitive attributes do not influence the outcomes unfairly. Using this intuition, we extend it by not only considering individual-level outcomes but also including a group-level distribution of historical interactions for a fairness reference point.

3.3 Proposed Fair Re-ranking Idea

When designing our re-ranking approach, we want to adjust the recommendations based on how users of different sensitive groups interact with items of different categories. For this, we leverage the popularity of different categories among users with different sensitive attributes from their historical interactions. By doing so, we effectively simulate a counterfactual scenario, where we use category preferences of users who do not share the same sensitive attributes. To achieve this we want to ensure that the deviation between:

- the category distribution recommended to a user, where the proportion of a single category is defined by $r(c|u, I)$ (as shown in Eq. 2) and
- the average category distribution of users who don't share the same sensitive attribute as this user, where the proportion of a single category is defined by $o(c|s_u)$ (as shown in Eq. 1) is minimized.

Essentially, $o(c|s_u)$ acts as a counterfactual baseline for us that helps counteract the tendency of recommenders to reinforce existing biases from the data they are trained on. This intuition will help align users' recommendations and act as a defying mechanism for historical bias.

To quantify the disparity between the two distributions, we will use KL divergence. Using Kullback-Leibler (KL) divergence, in this case, has numerous advantages, such as sensitivity to subtle differences in the two category distributions, alignment with counterfactual definition, where we capture the difference in how recommended items differ when sensitive attributes are changed and ease of interpretation. The equation below helps quantify the disparity between the two distributions:

$$D_{KL}(o||r(I)|u) = \sum_{c \in \mathcal{C}} o(c|s_u) log \frac{o(c|s_u)}{\tilde{r}(c|u, I)} \tag{3}$$

where

$$\tilde{r}(c|u, I) = (1 - \alpha) \cdot r(c|u, I) + \alpha \cdot o(c|s_u)$$

To avoid getting any value of $r(c|u, I) = 0$, we use $\tilde{r}$ where α is a really small number between 0 and 1. Note that here, o and $r(I)$ represent the distribution of CCP and RCP across all categories for user u.

We use an adaptation of Maximum Marginal Relevance (MMR) [7] to determine the optimal set of items I^*, which can be formalized as:

$$I^* = \text{argmax}_{I \subseteq TopN, |I|=k} (1 - \beta) \cdot rel(I, u) - \beta \cdot D_{\text{KL}}(o, r(I), u) \tag{4}$$

where

$$rel(I,u) = \sum_{v \in I} \text{score}_{u,v}$$

We use a hyperparameter $\beta \in [0,1]$ to calibrate the trade-off between relevance and fairness like some previous works, including [25,35,48]. This gives us a combinatorial optimization problem that is NP-hard. Following the works by [46,48], which demonstrated that the greedy optimization of an equation similar to Eq. 4 is equivalent to the greedy optimization of a surrogate submodular function, we adopt a similar approach condensing our equation to:

$$I^* = \text{argmax}_{I \subseteq TopN, |I|=k}(1-\beta) \cdot rel(I,u) + \beta \cdot \sum_c o(c|s_u) log \sum_{j=1}^{|I|} \frac{1}{j^\gamma} \tilde{r}(c|v_j) \quad (5)$$

where $\tilde{r}(c|v_j) = (1-\alpha) \cdot r(c|v_j) + \alpha \cdot o(c|s_u)$, and represents the proportion of category c in movie v_j. The simplified submodular greedy optimization has an optimal guarantee of $1 - \frac{1}{e}$ [41]. The algorithm for this optimization is presented as Algorithm 1. Here we generate the top N items for each user (represented by $TopN$) and then re-rank to find the top k items (where $N \geq k$). Instead of using $o(c|s_u)$ directly, we add a small constant variation across all c values (the impact of which can be considered negligible) to ensure non-zero entries. Additionally, we normalize the relevance term and fairness term through the min-max normalization scheme to ensure they are on the same scale.

We want to mention that although we aim to provide fair recommendations to the users based on their sensitive attributes, we ensure this does not come at the expense of personalization. For our fair scheme, the goal is still prioritizing the preferences of users, but in a way that prevents the reinforcement of social stereotypes.

4 Experimental Methodology

4.1 Datasets

We evaluate the effectiveness of our scheme on three publicly available datasets from different domains as shown in Table 2. The datasets are all pre-processed to remove items and users by k-core filtering, which is a common practice adopted in prior research [1,10,28]. In our case, we use 5-core filtering. For the Yelp dataset, we follow Kheya et al. [28] and condense the number of categories from over 300 to 21.

4.2 Baselines

As suggested by [14], we evaluate our re-ranking scheme on several recommendation approaches, including traditional ones (Biased Matrix Factorization [29] and Weighted Matrix Factorization [21,43]) and deep learning-based ones (Neural Matrix Factorization [20] and Variational Auto Encoder Collaborative Filtering

Algorithm 1. The Counterfactually Fair Re-ranking Optimization

Input: $\mathcal{U}, TopN, \beta, k, scores, S$

Output: Matrix R of size $|\mathcal{U}| \times k$ which contains fair top-k recommendation lists for each user.

$scores \leftarrow$ train baseline model and store scores of candidate items.
$R \leftarrow$ empty matrix of size $|\mathcal{U}| \times k$
$H(u) \leftarrow$ store historical interactions of all users.
Compute o for all possible s_u values, following Equation 4 for chosen S
for all users $u \in \mathcal{U}$ **do**
  for index = 0 **to** k-1 **do**
    for all items $i \in TopN_u \setminus R(u)$ **do**
      Compute fairness-aware scores for i using:

$$(1-\beta) \cdot rel(I,u) + \beta \cdot \sum_{c} o(c|s_u) \log \sum_{j=1}^{|R(u)\cup i|} \frac{1}{j^{\gamma}} \tilde{r}(c|v_j)$$

    end for
    Select the item i^* with the highest fairness-aware score.
    Add i^* to $R(u)$.
  end for
end for
return R

Table 2. Details of the three datasets along with the sensitive attributes, where G=Gender, A=Age, and O=Occupation. The number after each sensitive attribute represents the number of classes for that sensitive attribute. For instance, G: 2 means gender has two classes-[male, female]. Note: in our experiments, we use binary gender, but our method can be applied to non-binary genders as well.

Name	Interactions	Users	Items	Sensitive Attribute	Categories
ML-100K [19]	99,278	943	1,348	G: 2, A: 7, O: 21	18
ML-1M [19]	999,611	6,040	3,416	G: 2, A: 7, O: 21	18
Yelp [37]	97,991	1,316	1,272	G: 2	21

[33]). For all the models, we choose the best one based on the HitRatio@20 and NDCG@20 values after running them over multiple epochs for different combinations of hyperparameters. We empirically discovered that for weighing ranked items, using a gamma value of 0.1 works best in both reducing bias and minimizing performance degradation (refer to Fig. 2). For the smaller datasets, we use N as the total number of items in the dataset. For the 1M datatset, N is chosen to be 1000, and $TopN_u$ for each user is the top 1,000 items for user u.

Bias. For calculating bias, we extend two metrics introduced by [28], which take into account the distribution of categories of items recommended. They were originally used to quantify gender bias in recommendations. We extend them and use them to find the sum of pair-wise differences between all user

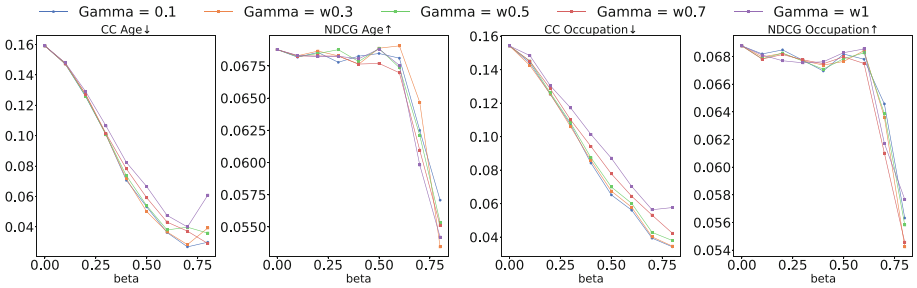

Fig. 2. Impact of γ across different β values on NDCG and CC values for VAE-CF model for ML100K.

groups, as suggested by the authors, to quantify bias in multi-valued sensitive attribute groups. The first metric calculates disparity in category distributions without considering the rank, like so:

$$CC(c,\mathcal{U}) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{|TopK_u|}\sum_{v\in TopK_u}\frac{C_{v,c}}{|C_v|} \tag{6}$$

The second metric scores items by discounting them based on the rank of the items in the top k list. This equation is formalized as:

$$CDCG(c,\mathcal{U}) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{|TopK_u|}\sum_{j=1}^{|TopK_u|}\frac{\frac{C_{v_j,c}}{|C_{v_j}|}}{\log(j+1)} \tag{7}$$

The sum of pairwise differences in CC and $CDCG$ values are then summed across all categories to represent the final bias values.

Performance. To evaluate the performance of the models, we use HitRatio@k, which measures the proportion of users who get at least one relevant item recommended to them. Additionally, we use a ranking-based metric called NDCG@k (Normalized Cumulative Gain) to measure the quality of the recommendations by giving higher importance to relevant items appearing higher in the list. For all our calculations, we use $k = 20$.

5 Results

We present the results of our experiments in Fig. 3, Fig. 5, Fig. 4 and Table 3.

5.1 Baseline Comparison

Out of all the models, NeuMF is more prone to capturing bias relating to sensitive attributes of users. The underlying architecture of this model uses Generalized

Table 3. Performance and bias values across all three datasets for sensitive attributes Age (A), Gender (G), and Occupation (O).

ML100K

	MF								WMF							
	NDCG↑		HitRatio↑		CC↓		CDCG↓		NDCG↑		HitRatio↑		CC↓		CDCG↓	
	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orignal	Fair
A		0.0423		0.4210	0.1254	0.0297	0.0485	0.0107		0.0490		0.4708	0.2055	0.0302	0.0745	0.0109
G	0.0397	0.0405	0.3924	0.4019	0.1300	0.0624	0.0507	0.0166	0.0372	0.0405	0.3924	0.4210	0.1888	0.0409	0.0736	0.0131
O		0.0430		0.4231	0.1688	0.0335	0.0650	0.0121		0.0483		0.4719	0.2248	0.0331	0.0830	0.0112
	VAE-CF								**NeuMF**							
	NDCG↑		HitRatio↑		CC↓		CDCG↓		NDCG↑		HitRatio↑		CC↓		CDCG↓	
	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orignal	Fair
A		0.0681		0.5451	0.1591	0.0362	0.0601	0.0135		0.0688		0.5758	0.2450	0.0295	0.0877	0.0107
G	0.0688	0.0682	0.5387	0.5419	0.0516	0.0371	0.0214	0.0177	0.0708	0.0688	0.5440	0.5567	0.2275	0.0253	0.0850	0.0099
O		0.0678		0.5355	0.1542	0.0561	0.0574	0.0200		0.0717		0.5769	0.2402	0.0358	0.0878	0.0136

ML1M

	MF								WMF							
	NDCG↑		HitRatio↑		CC↓		CDCG↓		NDCG↑		HitRatio↑		CC↓		CDCG↓	
	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orignal	Fair
A		0.0378		0.3768	0.1392	0.0537	0.0509	0.0202		0.0465		0.4722	0.2228	0.0217	0.0802	0.0080
G	0.0363	0.0377	0.3474	0.3700	0.2016	0.0629	0.0708	0.0261	0.0428	0.0455	0.4394	0.4662	0.3643	0.0447	0.1334	0.0144
O		0.0376		0.3732	0.1286	0.0597	0.0476	0.0224		0.0464		0.4727	0.2374	0.0234	0.0867	0.0103
	VAE-CF								**NeuMF**							
	NDCG↑		HitRatio↑		CC↓		CDCG↓		NDCG↑		HitRatio↑		CC↓		CDCG↓	
	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orignal	Fair
A		0.0510		0.4985	0.2076	0.0330	0.0758	0.0149		0.0470		0.4778	0.2554	0.0219	0.0928	0.0081
G	0.0515	0.0513	0.4616	0.4884	0.2603	0.0667	0.0969	0.0309	0.0478	0.0451	0.4389	0.4684	0.4101	0.0593	0.1484	0.0159
O		0.0519		0.4959	0.1514	0.0677	0.0542	0.0274		0.0468		0.4828	0.2562	0.0148	0.0928	0.0068

Yelp

	MF								WMF							
	NDCG↑		HitRatio↑		CC↓		CDCG↓		NDCG↑		HitRatio↑		CC↓		CDCG↓	
	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orignal	Fair
G	0.0202	0.0203	0.2660	0.2690	0.0158	0.0202	0.0072	0.0062	0.0264	0.0258	0.3590	0.3389	0.0502	0.0247	0.0178	0.0088
	VAE-CF								**NeuMF**							
	NDCG↑		HitRatio↑		CC↓		CDCG↓		NDCG↑		HitRatio↑		CC		CDCG	
	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orig.	Fair	Orignal	Fair
G	0.0900	0.0840	0.7196	0.6877	0.0670	0.0261	0.0235	0.0080	0.0897	0.0837	0.7310	0.6960	0.0617	0.0245	0.0222	0.0065

Matrix Factorization (GMF) and Multi-Layer Perceptrons (MLP), that captures the relationships between users and the items they have interacted with. This can cause the model to capture intricate details about user preferences, which can reflect societal stereotypes [28]. VAE-CF uses probabilistic variational auto-encoders to learn user-item interactions by encoding them in latent space. While it is sensitive to capturing biases, as seen from Table 3, the effect is minimal when compared to the MF-based models. Across all the models, the bias scores are higher for the larger dataset, likely due to the fact that more interactions help provide more opportunities for the model to capture the underlying biases. For performance, the deep model performs better in almost all cases, which is expected. Gender-related bias is more pronounced across all models. We believe age and occupation, may have a more subtle impact on category preferences when compared to gender. Since age and occupation have more classes than gender, the bias is more spread out across these groups making the impact more diluted. Imbalances in the interactions when considering just binary gender will stand out more, making gender bias more evident.

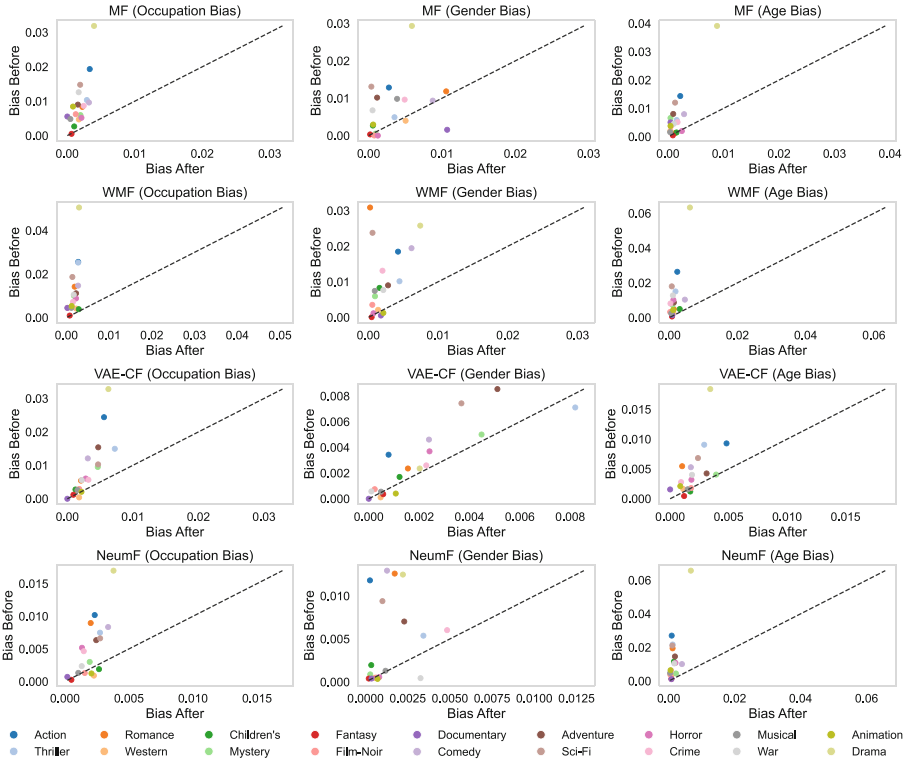

Fig. 3. Comparison of bias values (CC) before and after fair re-ranking for all the models across all categories. This is only visualized for the ML100K dataset; however, other datasets have similar trends. We plot CC scores because CC and CDCG are correlated and exhibit similar patterns.

5.2 Impact of Fair-Reranking

Bias. There is significant bias reduction after applying our re-ranking scheme for all models across the three sensitive attributes, as observed in Fig. 3, Fig. 4 and Table 3. From Fig. 3, we can observe how most points are above the y=x diagonal, verifying the reliable effectiveness of the re-ranking scheme in reducing bias. The bias mitigation works best for the NeuMF model since there's a noticeable drop in both CC and CDCG values across all the datasets for all three sensitive attributes. In Fig. 4, the effectiveness of our re-ranking approach is evident in plots c and e where the disparities in category distributions of recommended movies are significantly reduced compared to the baseline models (plots b and d). We also emphasize how our re-ranking approach achieves fairness without compromising the preferences of users. For instance, both genders have strong preferences for comedies as seen in plot a, but this is not reflected in the recommendations from the VAE-CF model (plot d). However, in our fair model (plot e), the proportion of comedy movies recommended is aligned with that

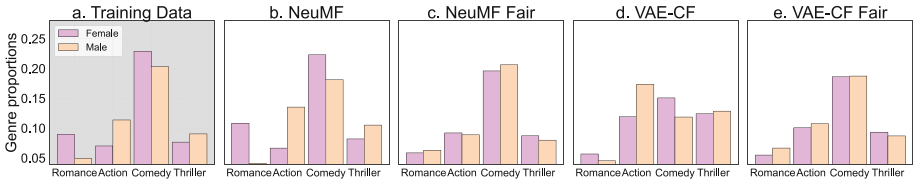

Fig. 4. Comparison of recommendations before and after fair re-ranking for two of the best-performing models for the ML1M dataset across four stereotypical genres. We also show the genre proportions of the training dataset.

of the training set. The proportions for each category are calculated following the CC formula. In the case of the training set we use the historical interactions instead of the top-k recommendations. While, Fig. 4, only shows results for ML1M dataset, readers can refer to Fig. 1, for ML100K dataset results. We did not include visualizations of the Yelp dataset, since they follow similar drops in bias values, like those of the other datasets. Our approach effectively minimizes discrepancies in recommended restaurants of different categories like *Coffee, Tea & Desserts* (which is more biased towards female users) and *Travel & Transportation* (which is more biased towards male users).

In most cases, the performance metrics are observed to be increasing while the bias is mitigated. Theoretically, as β increases, we would expect a decrease in bias scores and NDCG value. After a certain value of β, we would expect the bias scores to increase since the re-ranking algorithm would essentially allow the bias from CCP distribution to dominate over the actual bias. This would, in turn, start increasing the bias in the opposite direction (although we don't consider the direction of the bias because we use absolute values, we mention it here for clarity in explaining the phenomenon). To observe the influence of β on recommendation performance and bias, we run our re-ranking algorithm for all models for β values from 0 to 0.8, with increments of 0.1. We don't include values above 0.8 since it doesn't make sense to over-power the actual relevance scores. As seen in Fig. 5, there is a general decrease in bias values over the first few values of β. For gender, the bias starts increasing after β is greater than a certain value (different for different models). While this trend is more prominent in the case of gender, it is also observed for age. Our intuition for a profound bias increase in gender is that the CCP distribution we are employing to fix the bias is stronger in the case of gender. So, while the distribution helps us reduce bias, as β increases, the bias increases in the opposite direction. Again, age and occupation have more classes, and the average bias from all of these is too diluted to impact too strongly when we are using them to mitigate bias of each user's recommendations. For most models, a β value close to 0.4-0.6, seems to work well for mitigating bias.

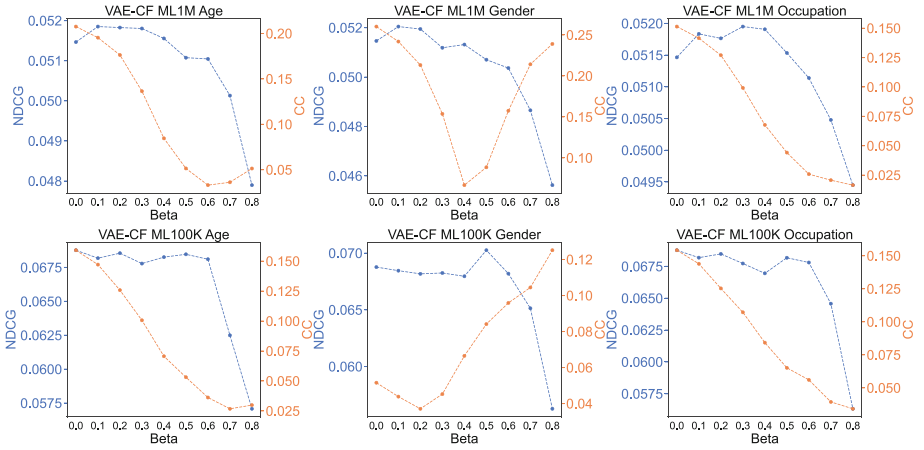

Fig. 5. Impact of β for VAE-CF model. The other models follow similar trends.

Performance. From Fig. 5, we can observe a trend where the performance increases slightly and then decreases. While the increase in performance seems counterintuitive, this can be because the fairness term helps reduce overfitting. One way to think about this, is how the fairness term is indirectly improving the coverage of the recommendations, which in turn provides relevant items to the users. Since our approach improves the exposure of items across categories, it enhances user engagement. Additionally, the idea of improving fairness leading to improvement in performance has been observed previously by [23,26,28,40]. While there are some instances where there is performance drop due to increased fairness as observed in Table 3, the decline is kept to a minimum.

6 Conclusion

In this work, we recognize the underlying issues of the current re-ranking approaches to mitigate bias in recommendations. We introduce a re-ranking scheme that reliably mitigates social bias for multi-valued user-sensitive attributes while also using item categories to ensure fine-grained treatment. Our approach is a simple yet powerful post-processing scheme to mitigate bias, which requires no modification of the model's internal parameters. We show, through extensive experiments, on three real-world datasets from multiple domains, the effectiveness of our re-ranking approach. The results show how our approach not only helps reduce bias but also preserves the quality of the recommendations, with a negligible drop in performance. We leverage the bias in the dataset to correct biased recommendations. While this works well for currently used datasets since they have historical bias, it may be less useful if future datasets evolve to be more neutral and unbiased. But we believe a dataset without bias (unless

explicitly preprocessed to be fair) remains a distant possibility. While this work mainly focuses on consumers, it also implicitly accounts for the provider-side since we include item categories. In the future, we want to explicitly address the provider perspective (for instance, including item brands) to ensure a more holistic solution to social bias in recommendations. Additionally, we also intend to extend our work to address intersectional fairness for the consumers (like female and doctor). Our code, along with the processed datasets, are available here: Re-ranking Code[1]

Acknowledgment. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA2386-23-1-4003.

References

1. Anelli, V.W., et al.: Elliot: a comprehensive and rigorous framework for reproducible recommender systems evaluation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021, pp. 2405–2414. ACM, New York, NY, USA (2021)
2. Beutel, A., et al.: Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, pp. 2212–2220. (2019)
3. Bhadani, S.: Biases in recommendation system. In: Proceedings of the 15th ACM Conference on Recommender Systems, RecSys 2021, pp. 855–859. (2021)
4. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS 2016, pp. 4356–4364 (2016)
5. Boratto, L., Fenu, G., Marras, M.: Interplay between upsampling and regularization for provider fairness in recommender systems. User Model. User-Adap. Inter. **31**(3), 421–455 (2021)
6. Burke, R.: Multisided Fairness for Recommendation (2017), arXiv:1707.00093
7. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 335–336 (1998)
8. Datta, A., Tschantz, M., Datta, A.: Automated experiments on ad privacy settings. In: Proceedings on Privacy Enhancing Technologies, vol. 1 (2015)
9. De-Arteaga, M., et al.: Bias in bios: a case study of semantic representation bias in a high-stakes setting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 120–128 (2019)
10. Dietz, L.W., Sánchez, P., Bellogín, A.: Understanding the influence of data characteristics on the performance of point-of-interest recommendation algorithms. Inf. Technol. Tourism **27**(1), 75–124 (2025)
11. Edizel, B., Bonchi, F., Hajian, S., Panisson, A., Tassa, T.: Fairecsys: mitigating algorithmic bias in recommender systems. Int. J. Data Sci. Analytics **9**, 197–213 (2020)

[1] https://github.com/tahsinkheya/re_ranking_clean.

12. Ekstrand, M.D., et al.: All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 172–186 (2018)
13. Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C., Venkatasubramanian, S.: Runaway feedback loops in predictive policing. In: Friedler, S.A., Wilson, C. (eds.) Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Proceedings of Machine Learning Research, vol. 81, pp. 160–171. PMLR (2018)
14. Ferrari Dacrema, M., Cremonesi, P., Jannach, D.: Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, pp. 101–109 (2019)
15. Fu, Z., et al.: Fairness-aware explainable recommendation over knowledge graphs (2020). https://doi.org/10.48550/arXiv.2006.02046
16. Geyik, S.C., Ambler, S., Kenthapadi, K.: Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, pp. 2221–2231 (2019)
17. Gorantla, S., Deshpande, A., Louis, A.: On the problem of underranking in group-fair ranking. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 3777–3787 (2021)
18. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
19. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. ACM Trans. Interact. Intell. Syst. **5**(4) (2015)
20. He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.S.: Neural collaborative filtering. In: Proceedings of the 26th International Conference on World Wide Web, pp. 173–182. International World Wide Web Conferences Steering Committee (2017)
21. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 263–272 (2008)
22. Hua, W., Ge, Y., Xu, S., Ji, J., Li, Z., Zhang, Y.: UP5: unbiased foundation model for fairness-aware recommendation. In: Graham, Y., Purver, M. (eds.) Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1899–1912. ACL, St. Julian's, Malta, March 2024
23. Islam, R., Keya, K.N., Zeng, Z., Pan, S., Foulds, J.: Debiasing career recommendations with neural fair collaborative filtering. In: Proceedings of the Web Conference 2021, pp. 3779–3790 (2021)
24. Jaenich, T., McDonald, G., Ounis, I.: Fairness-aware exposure allocation via adaptive reranking. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1504–1513 (2024)
25. Karako, C., Manggala, P.: Using image fairness representations in diversity-based re-ranking for recommendations. In: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, pp. 23–28 (2018)
26. Keya, K.N., Islam, R., Pan, S., Stockwell, I., Foulds, J.R.: Equitable allocation of healthcare resources with fair cox models (2020), https://arxiv.org/abs/2010.06820

27. Kheya, T.A., Bouadjenek, M.R., Aryal, S.: The pursuit of fairness in artificial intelligence models: a survey (2024), https://arxiv.org/abs/2403.17333
28. Kheya, T.A., Bouadjenek, M.R., Aryal, S.: Unmasking gender bias in recommendation systems and enhancing category-aware fairness. In: Proceedings of the ACM Web Conference 2025 (2025)
29. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**(8), 30–37 (2009)
30. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 30 (2017)
31. Li, Y., Chen, H., Fu, Z., Ge, Y., Zhang, Y.: User-oriented fairness in recommendation. In: Proceedings of the Web Conference 2021, pp. 624–632 (2021)
32. Li, Y., Chen, H., Xu, S., Ge, Y., Zhang, Y.: Towards personalized fairness based on causal notion. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2021, pp. 1054–1063 (2021)
33. Liang, D., Krishnan, R.G., Hoffman, M.D., Jebara, T.: Variational autoencoders for collaborative filtering. In: Proceedings of the 2018 World Wide Web Conference, pp. 689–698 (2018)
34. Lin, K., Sonboli, N., Mobasher, B., Burke, R.: Crank up the volume: preference bias amplification in collaborative recommendation (2019), https://arxiv.org/abs/1909.06362
35. Liu, W., Guo, J., Sonboli, N., Burke, R., Zhang, S.: Personalized fairness-aware re-ranking for microlending. In: Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, pp. 467–471 (2019)
36. Lum, K., Isaac, W.: To predict and serve? Significance **13**, 14–19 (2016)
37. Mansoury, M., Mobasher, B., Burke, R., Pechenizkiy, M.: Bias disparity in collaborative recommendation: algorithmic evaluation and comparison (2019), https://arxiv.org/abs/1908.00831
38. Mehrotra, A., Vishnoi, N.: Fair ranking with noisy protected attributes. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems, vol. 35, pp. 31711–31725. Curran Associates, Inc. (2022)
39. Melchiorre, A.B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., Schedl, M.: Investigating gender fairness of recommendation algorithms in the music domain. Inf. Process. Manag. **58**(5), 102666 (2021)
40. Naghiaei, M., Rahmani, H.A., Deldjoo, Y.: Cpfair: personalized consumer and producer fairness re-ranking for recommender systems. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 770–779 (2022)
41. Nemhauser, G.L., Wolsey, L.A., Fisher, M.L.: An analysis of approximations for maximizing submodular set functions–i. Math. Program. **14**, 265–294 (1978)
42. Neophytou, N., Mitra, B., Stinson, C.: Revisiting popularity and demographic biases in recommender evaluation and effectiveness. In: European Conference on Information Retrieval. BCS-IRSG, Springer (2021)
43. Pan, R., Zhou, Y., Cao, B., Liu, N.N., Lukose, R., Scholz, M., Yang, Q.: One-class collaborative filtering. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 502–511 (2008)
44. Patro, G.K., Porcaro, L., Mitchell, L., Zhang, Q., Zehlike, M., Garg, N.: Fair ranking: a critical review, challenges, and future directions. In: Proceedings of the 2022

ACM Conference on Fairness, Accountability, and Transparency, FAccT 2022, pp. 1929–1942 (2022)
45. Rastegarpanah, B., Gummadi, K.P., Crovella, M.: Fighting fire with fire: using antidote data to improve polarization and fairness of recommender systems. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, pp. 231–239 (2019)
46. Shinohara, Y.: A submodular optimization approach to sentence set selection. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4112–4115 (2014)
47. Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2219–2228 (2018)
48. Steck, H.: Calibrated recommendations. In: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, pp. 154–162 (2018)
49. Sweeney, L.: Discrimination in online ad delivery. Commun. ACM **56**(5), 44–54 (2013)
50. Thonet, T., Renders, J.M.: Multi-grouping robust fair ranking. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, pp. 2077–2080. ACM, New York, NY, USA (2020)
51. Wang, Y., Ma, W., Zhang, M., Liu, Y., Ma, S.: A survey on the fairness of recommender systems. ACM Trans. Inf. Syst. **41**(3), 1–43 (2023)
52. Wei, T., He, J.: Comprehensive fair meta-learned recommender system. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2022, pp. 1989–1999 (2022)
53. Wu, C., Wu, F., Wang, X., Huang, Y., Xie, X.: Fairness-aware news recommendation with decomposed adversarial learning. Proc. AAAI Conf. Artif. Intell. **35**(5), 4462–4469 (2021)
54. Wu, L., Chen, L., Shao, P., Hong, R., Wang, X., Wang, M.: Learning fair representations for recommendation: a graph-based perspective. In: Proceedings of the Web Conference 2021, pp. 2198–2208 (2021)
55. Yang, K., Loftus, J.R., Stoyanovich, J.: Causal intersectionality for fair ranking. CoRR **abs/2006.08688** (2020), https://arxiv.org/abs/2006.08688
56. Yao, S., Huang, B.: Beyond parity: fairness objectives for collaborative filtering. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
57. Yesilada, M., Lewandowsky, S.: Systematic review: Youtube recommendations and problematic content. Internet Policy Rev. **11**(1) (2022)
58. Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., Baeza-Yates, R.: Fa*ir: a fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (2017)