The Effect of Gender De-biased Recommendations — A User Study on Gender-specific Preferences

Thorsten Krause*
German Research Center for Artificial Intelligence
Osnabrück, Germany
Radboud University
Nijmegen, Netherlands
thorsten.krause@ru.nl

Lorena Göritz
German Research Center for Artificial Intelligence
Osnabrück, Germany
lorena.goeritz@dfki.de

Robin Gratz
German Research Center for Artificial Intelligence
Osnabrück, Germany
robin.gratz@dfki.de

Abstract

Recommender systems treat users inherently differently. Sometimes, however, personalization turns into discrimination. Gender bias occurs when a system treats users differently based on gender. While most research discusses measures and countermeasures for gender bias, one recent study explored whether users *enjoy* gender de-biased recommendations. However, its methodology has significant shortcomings; It fails to validate its de-biasing method appropriately and compares biased and unbiased models that differ in key properties. We reproduce the study in a 2x2 between-subjects design with $n = 800$ participants. Moreover, we examine the authors' hypothesis that educating users on gender bias improves their attitude towards de-biasing. We find that the genders perceive de-biasing differently. The female users —the majority group— rate biased recommendations significantly higher while the male users —the minority group— indicate no preference. Educating users on gender bias increased acceptance non-significantly. We consider our contribution vital towards understanding how gender de-biasing affects different user groups.

CCS Concepts

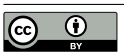
• **Information systems** → **Recommender systems**; • **Human-centered computing** → **Empirical studies in accessibility**; *Accessibility design and evaluation methods*; *Accessibility technologies.*

Keywords

Gender Bias, Recommender Systems, Fairness, User Study, Reproducibility

ACM Reference Format:
Thorsten Krause, Lorena Göritz, and Robin Gratz. 2025. The Effect of Gender De-biased Recommendations — A User Study on Gender-specific Preferences. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3706598.3713155

*Both authors contributed equally to this research.



CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713155

1 Introduction

Humans interact with recommender systems in nearly all areas of life, from news [13, 63], social media [52], to entertainment [12, 15, 41]. The design of this interaction, however, goes beyond recommendation accuracy [8]. For example, how many items to recommend [6], how to organize them [37], and how to present them [1, 4] are all design decisions that can influence the user-system interaction.

In recent years, the design of the recommendation algorithms caught attention for its role in the user-model feedback loop [10, 26, 33]. When the system gives recommendations, the user provides feedback, for example, through clicks or ratings and the system updates its user model based on their feedback. Because the system determines which information the user receives, it has the potential to shape the user's opinion and behavior over time [39]. Consequently, poorly designed algorithms can lead to propagation and reinforcement of biases within the user-model feedback loop [10].

One harmful bias prevalent in recommender systems is gender bias [20, 50]. Gender bias occurs when the model does not treat all genders equally and is especially critical in the domain of educational or career recommendation. Existing stereotypes in the training data, such as females being statistically associated with health care-related occupations, can be learned and amplified by recommendation models, leading to systematically different recommendations for different genders [60]. For example, Rus et al. [47] find that gender bias leads to women being recommended jobs with significantly lower salaries than men. Another example is the now well-known incident at Amazon where an experimental job candidate recommendation tool systematically scored women's resumes lower than men's [14].

A plethora of research on fairness and bias mitigation in recommender systems exists [34, 44, 47, 64]. However, previous studies only showed that their proposed de-biasing methods return fairer recommendations. To decide whether to implement those de-biasing methods, practitioners need to know how users react to the de-biased recommendations. On the one hand, users could prefer the de-biased recommendations because they overcome traditional stereotypes. On the other hand, users could prefer the biased recommendations because they align with their existing beliefs and expectations. In some settings, a decrease in satisfaction for one group could be acceptable if it leads to an increase in satisfaction for another, disadvantaged group. According to the authors, and to our own knowledge, Wang et al. [60] were the first to pursue this question. In the context of college major recommendation, they

constructed a biased and a de-biased model and let 200 college students rate its output. They found that participants preferred the biased over the gender de-biased recommendations.

The study, however, has important limitations. First, the offline evaluation that was supposed to confirm the de-biasing method's functionality prior to the user study was designed in favor of the de-biased model: The study compared the biased with the de-biased model on an existing dataset and found that the de-biased model returned fairer results at higher accuracy. This result is surprising since de-biasing relates to a constrained problem with a generally worse optimal solution than the unconstrained (biased) problem. Indeed, to our understanding, the de-biased model had access to more training data than the biased model. Hence, we do not know how well the de-biased model would perform in a fair comparison. However, we need this information for interpreting the user study's results; only if we know that the de-biasing retains performance can we attribute changes in user satisfaction to the increase in fairness. Second, the de-biased model's architecture differed from the biased model's, which renders a direct comparison between the biased and de-biased recommendations impossible. Specifically, the de-biased model deliberately missed an intercept term to *"further remove popularity bias"* [60]. However, by de-biasing with respect to both gender and popularity bias at the same time, the study's results cannot be clearly attributed to gender de-biasing. Instead, the observed effects could be caused by the popularity de-biasing. Third, fairness was measured through the *Non-parity unfairness* (U_{Par}) measure [65], which we argue is not sufficient to conclude on the method's de-biasing capabilities in Section 3. Hence, an underperforming gender de-biasing method could have compromised the study's results. Indeed, other de-biasing methods dramatically out-performed the de-biasing method from [60] in our study. Accordingly, we derived our first research question:

RQ1 *Can we reproduce the findings of Wang et al. [60] with verified de-biasing methods and comparable models?*

Wang et al. [60] followed their experiment up with a questionnaire on how acceptance towards de-biased recommendations could be improved. Participants indicated that their trust in a fair recommender system could be improved through providing additional information on what gender bias is, its emergence, and how it affects societal gender stereotypes and fairness. We investigate whether such information benefit users' perception of the system to answer the following research question:

RQ2 *How do explanations on gender bias influence user perception of gender de-biased educational recommendations?*

Last, the study only measured the accumulated effects over both genders. We hypothesize that male and female users perceive gender de-biased recommendations differently, especially when one group is a minority. Hence, we define our third research questions:

RQ3 *How does gender influence user perception of gender de-biased educational recommendations?*

We conduct an online user study with a 2x2 between-subjects design in which $n = 800$ participants evaluate recommendations for educational courses. Participants first indicate their course preferences among a catalog of non-job-specific educational courses for adults. Then, they view their personalized course recommendations and indicate their satisfaction. Half of all participants receive recommendations from a biased systems, half from a de-biased systems to investigate **RQ1**. To examine **RQ2**, half of all users receive additional information on gender bias in recommender systems and its implications before receiving their recommendations. We analyze the difference in the results by gender to address **RQ3**.

Contrary to the original study [60], we thoroughly assess several candidate de-biasing methods, identify a more potent de-biasing method, and group the results by gender to investigate potential differences between the two groups. We discuss and demonstrate the limitations of the U_{PAR} fairness measure and address them by proposing the *de-biasing correlation coefficient (DCC)* as a novel, complementary metric. Moreover, we propose *gender deconfounding (GD)*, an causal de-biasing inspired method. GD displays significantly stronger de-biasing capabilities according to both U_{PAR} and DCC than the de-biasing method from the original study.

Our study adds to the young body of research on the user-side of gender de-biased recommendations. Specifically, we aim to understand how gender de-biasing as a design element [18], as well as providing the user with additional context on gender bias, affects the user-system interaction. In summary, our study's main contributions are:

- We reproduce the study by Wang et al. [60] while overcoming its shortcomings (**RQ1**).
- We test whether explanations on gender bias can improve user perception of gender de-biased recommendations (**RQ2**).
- We show that different genders benefit differently from gender de-biasing (**RQ3**).
- We propose the DCC for measuring the effect of gender de-biasing on rankings and demonstrate its advantage over U_{PAR}, which was used in the original study.
- We propose a causal gender de-biasing method, GD, using gender-specific item bias factors.

The remainder of this work is structured as follows: Section 2 addresses the theoretical background and gives information on recommender systems, gender bias, and bias mitigation. Section 3 defines and discusses the de-biasing measures and methods that are relevant for the following sections. Section 4 describes our online user study and survey. Section 5 summarizes our results, which we discuss with respect to our research questions in Chapter 6. Lastly, we conclude in Section 7.

2 Related Work

Prior to our study, we performed a structured literature search on the intersection of gender bias and recommender systems. After comparing the search results of three different queries, we retrieved all 1041 results for the query *"gender bias" AND "recommender"*. After screening titles and filtering those that included references to either gender, recommender systems, AI ethics, or AI fairness, (254 papers), we filtered any papers whose abstract related to gender or and recommender systems (67 papers). Further, we excluded any inaccessible papers and any papers that specifically related to the topic of recommendation letters or multi-sided fairness in recommenders, although the latter was later re-included. We also discarded any books, theses and non-peer-reviewed papers. Then,

we screened the full texts of the remaining 63 studies and found that 24 were relevant to our research topic. As an exhaustive report and analysis of the identified literature are beyond the scope of this article, we provide an overview of the most relevant findings, enriched by a forward- and backward search [61] and complementary references on related topics.

2.1 Fairness in Recommendation

Many fairness concepts for recommendation exist and most imply equal treatment of stakeholders. The stakeholder-level is often divided into individuals and groups [34]. *Individual fairness* means that similar individuals get similar outcomes [60]. *Group fairness* means that groups receive similar outcomes [23]. Groups can be defined by (sensitive) attributes, e.g. gender [60]. Laterally, *inter-user (user-side) fairness* describes fairness among users and *inter-item (item-side) fairness* describes fairness among items [11]. However, the literature has vastly different understandings of what is fair. Wu et al. [64] cluster the most popular fairness concepts in recommendation as follows:

- *Equal Opportunity [22]:* A model should produce the same true-positive rate for all stakeholders.
- *Envy-freeness [16]:* Every stakeholder should prefer their recommendations over those of other stakeholders.
- *Demographic Parity [65]:* Decisions should be similar around sensitive attributes (e.g. gender).
- *Counterfactual Fairness [35]:* For any individual, any prediction in the real world should be the same as in a counterfactual world.
- *Fairness through Awareness [2]:* This definition is equivalent to the concept of individual fairness [64].
- *Fairness through Unawareness [63]:* The model should not know the protected attributes. Due to correlations with users' protected attributes and revealed preferences, this approach is considered ineffective [64].

Different fairness concepts can require different metrics and mitigation strategies. For example, *equal opportunity* is reached once the model performs comparably well for different stakeholders but recommendations can still differ between them. *Demographic parity* demands similar recommendations for different groups but does not imply equal performance.

2.2 Gender Bias

Glick and Fiske [21] find that sex is the primary, earliest learned, and most automatic category we use to classify a person. According to social role theory by Eagly and Wood [17], this observation originates from the physical differences between males and females. Besides men's greater size and strength, women's reproductive activities of pregnancy and lactation predispose them to nursing and taking care of children. West and Zimmerman [62] claim that the biological sex does not inherit gender, but that gender is something we do. *Doing* gender means shaping differences between women and men through societal interactions, rather than by natural or biological traits. They see gender as something that is achieved through certain "*socially guided perceptual, interactional, and micro-political activities that cast particular pursuits as expressions of masculine and feminine "natures"*" [62].

Gender stereotypes are, according to the European Commission, one main driver for inequality between men and women[1]. These stereotypes have a negative societal impact by confining individuals to rigid gender roles and preventing them from reaching their full potential. Education plays a critical role in addressing these challenges. Through lifelong learning, individuals develop new interests and refine their skills. However, participation in educational processes is often influenced by socialization into gender roles [57]. In order to provide equal opportunities for all, regardless of societal expectations, it is essential that educational recommendations provide comparable opportunities for men and women.

Information technology shapes our perception through the information that it provides us. Within the HCI-domain, previous studies examined gender bias in search engines, where results match user queries, but not in recommender systems where results match user preferences. Kay et al. [30] investigate gender bias in image search for different careers based on which genders are most prevalent in the results. They found that search engines exaggerate stereotypes, that people rate stereotypical search results better and that the representation of gender in search results affected people's perceptions about real-world distributions. We examine the second of the three aspects, expecting a similar finding based on the results by Wang et al. [60]. Otterbacher et al. [43] investigate search results for different character traits, finding that for some traits, people in the images were more likely to have a gender that is stereotypically associated to those traits. Kopeinik et al. [32] show that users themselves exhibit biases during the interaction with search engines by explicitly specifying gender in contra-stereotypical queries such as "show me a male nurse" and hypothesize that this behavior biases the systems, which can be expected as users who search for nurses would reward the system more for biased results. Aside from search engines, modern generative systems can incorporate gender biases from data into their output [46, 51] but not all do [48, 59]. That the interaction with technology can shape stereotypes was also confirmed for the case of computer games where assigning users to characters that embody marginalized groups can reduce their internal gender biases [25]. In the recommendation setting, where embodiment is not a viable design instrument, alternatives are needed.

In the recommendation context, not all research clearly defines and classifies gender bias. We found that the concepts usually relate to demographic parity or equal opportunity. When gender refers to the item-side, then gender bias entails group- and inter-item unfairness with respect to demographic parity. For example, Dacon and Liu [13] describe how news recommendation heavily influences and reinforces gender stereotypes. With newsworthy people (politicians, CEOs or athletes) mostly being men, women are underrepresented. Ferraro et al. [20] show how the user-model feedback loop can amplify gender bias against female artists. Mansoury et al. [40] find that on one dataset users' rating behavior, how informative their profile is, and the users' profile size, all of which correlate with a users' gender. When users receive different recommendations based on their gender, then gender bias entails group- and inter-user unfairness. Most studies consider demographic parity in

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0152

this context. For example, Melchiorre et al. [41] find that men receive more frequent recommendations of highly paid jobs, because women are expensive advertisement targets and the job recommendation algorithms minimize costs by targeting the less expensive males. In the online-user study by Wang et al. [60], male and female participants received different recommendations for college majors. The study argues that biased recommendations can lead to female students chose stereotypically female majors while male and female students should be exposed to similar recommendations. Mansoury et al. [40] considers gender bias with respect to equal opportunity and shows how attributes that correlate with female gender are also correlated with lower ranking performance on the MovieLens dataset.

Recommender systems can inherit gender bias from the training data [36]. Examining the gender wage gap, [47] find that gender can be predicted precisely, even from anonymized resumes. However, Ferraro et al. [19] find that the recommendation model has a greater effect on amplifying the bias than how users chose.

2.3 Gender Bias Mitigation

Wu et al. [64], [34] and Pitoura et al. [44] divide strategies to mitigate algorithmic bias in recommender systems into *(i) pre-processing*, *(ii) in-processing*, and *(iii) post-processing* methods. (i) Pre-processing techniques modify the training data. Most surveys refer to the classification in [28] into *suppression* [28], which removes sensitive attributes (e.g. gender) and correlated attributes, *massaging the dataset* [45], which manipulates the labels of some user interactions, *reweighing* the dataset [7], which balances the data without changing labels, and *over- and under-sampling* [41] which balances the distribution interactions related to majority and minority groups, again without changing the labels themselves. (ii) In-processing methods redesign the recommender algorithm. For example, one can *de-bias the latent factors and embeddings* [60] at the end of training or add a *fairness regularization term* [5, 65] or fairness constraint [66] to the objective function. Such a regularization term can be defined for different fairness notions or metrics, it could e.g. penalize a higher difference in demographic parity. (iii) Post-processing methods adjust existing rankings. *Generative strategies* [31] compute the fairness score of the recommended items and then generate a new ranking our of weighted relevance and fairness scores [65]. *Re-ranking* [29, 65] consider fairness metrics when computing interaction probabilities and then re-arrange the recommendations based on these values.

The only (i) pre-processing gender de-biasing technique that we are aware of is the one by Saxena and Jain [50], who aim to create fair exposure of female and male authors' books by cleaning the dataset of individual users' biased ratings. One (iii) post-processing technique was used by Ferraro et al. [20] who propose a re-ranking algorithm for increasing female artists' exposure. Apart from that, we find that most gender de-biasing methods are (ii) in-processing techniques: [47] apply adversarial de-biasing to generate user representations from which the protected attribute can no longer be recovered. *UGRec* [36] employs a similar strategy. FairRec [63] learns two separate user embeddings, a bias-aware representation that captures the bias information and a bias-free representation through adversarial learning and orthogonal regularization. [24] also investigate gender bias in career recommendation. To tackle the problem of sparse career information (users usually only have one or two college majors and a limited career path) they introduce Neural Fair Collaborative Filtering (NFCF). NFCF first trains the user and item embeddings on a regular NCF, then user embeddings are de-biased by orthogonal projection on the global bias vector.

Wang et al. [60], whose experiment our study replicates, employed the strategy by Islam et al. [24] in their case study in which they de-bias a recommender for college major recommendations and test it in an online experiment with 200 participants.

3 Relevant De-biasing Measures and Approaches

This section elaborates on the measures that we used to evaluate a recommendation algorithm's gender bias and the de-biasing methods that we evaluated as potential candidates for our study (Section 4). First, we discuss *Non-parity unfairness (U_{Par})* [65] that was used by Wang et al. [60] to evaluate their de-biasing method and propose a potential improvement. We then introduce and discuss a novel measure, the *De-biasing correlation coefficient (DCC)*, that evaluates the de-biasing method's *purity*. Afterwards, we define the base (biased) model architecture and three de-biasing methods including *Gender De-biasing (GD)*, a novel causal de-biasing method.

3.1 Measures for Bias and De-biasing

Non-parity unfairness (U_{Par}) [65]. The original study [60] compared the performance of their de-biasing method with U_{Par} [65]. Assume that the recommendation model bases the ranking on *relevance scores* U_{ij} for all users $i \in I$ and items $j \in J$. Most collaborative filtering models compute relevance scores to quantify a user's preference towards an item, and then sort the relevance scores in descending order to obtain the relevance *ranking*. The U_{Par} compares the expected predicted relevance scores U_{ij} between two user groups:

$$U_{\text{Par}} \coloneqq \frac{1}{|J|} \sum_{j \in J} \left| \mathbb{E}_{\text{female}} \left[U_{ij} \right] - \mathbb{E}_{\text{male}} \left[U_{ij} \right] \right| \tag{1}$$

A high U_{Par} score suggests that one group is more likely to receive recommendations for the item than the other. An effective de-biasing method should decrease the model's U_{Par} score.

However, the U_{Par} neglects the *ranking* that recommendations ultimately depend on. For example, scaling all relevance scores by the same positive constant $0 < c << 1$ reduces U_{Par} by a factor of c but does not affect the rankings. To address this limitation, we introduce $U_{\text{Par}}^{\text{rank}}$ which computes the U_{Par} score with respect to the *ranks* instead of the *relevance scores*. Furthermore, the U_{Par} score does not consider the de-biasing method's impact on recommendation accuracy. While accuracy metrics can measure the cost of a particular de-biasing method on the recommendation accuracy, they cannot reveal whether that cost is inherent to de-biasing or whether a better de-biasing method could exist. To address these limitations, we introduce a measure of the de-biasing method's *purity*, the *De-biasing correlation coefficient (DCC)*.

Non-parity rank unfairness ($U_{\text{Par}}^{\text{rank}}$). Based on the U_{Par}'s first limitation discussed above, that it depends on the relevance scores rather than the ranking, we propose the Non-parity rank unfairness

measure $U_{\text{Par}}^{\text{rank}}$ with

$$U_{\text{Par}}^{\text{rank}} := \frac{1}{|J|} \sum_{j \in J} \left| \mathbb{E}_{\text{female}} \left[r_{ij} \right] - \mathbb{E}_{\text{male}} \left[r_{ij} \right] \right|, \tag{2}$$

where r_{ij} is the rank of item j in the recommendation list for user i. Compared with U_{Par}, $U_{\text{Par}}^{\text{rank}}$ only reacts to changes in the rankings, not the scores per se. The interpretation and application of $U_{\text{Par}}^{\text{rank}}$ are analogous to U_{Par}. While both measures could be correlated in practice, we argue that $U_{\text{Par}}^{\text{rank}}$ more interpretable as it directly returns by how many ranks recommendations for an item differ between the two groups.

De-biasing correlation coefficient (DCC). Based on the U_{Par}'s second limitation, that it does reveal any possible side-effects of a de-biasing method, we propose the DCC. We first motivate and define the group-specific DCC and then derive the DCC.

The idea behind the DCC comes from an intuition about de-biasing. Demographic parity implies that courses should receive lower ranks for a user if they are systematically more preferred by their own group than by the opposite group. Assume that we know how many times any item was exposed to any group and how many times it was interacted with by that group. We define the ratio of both quantities, interactions by exposure, as the *choice ratio* for that group and that item. A de-biasing model should therefore rank those courses with a high male choice ratio and a low female choice ratios higher for female users and lower for male users than a biased model. Vice versa, it should rank those courses with a high female choice ratio and a low male choice ratio higher for male users and lower for female users. For example, in our study, the course *Vegan cooking - Black Forest Cherry Cake* had an 38% female choice ratio but only a 9% male choice ratio. A de-biasing model should reduce exposure to this course for female users and increase it for male users. Hence, the average difference in ranks δ_j^{rank} for female users between the biased and de-biased models should be correlated with the relative choice ratios of female and male users over all courses $j \in J$. The *group-specific de-biasing correlation coefficient* measures this correlation:

Definition 3.1. Given two disjoint groups $A, B \subset I$, let r_j^A and r_j^B be their respective relative choice ratios for items $j \in J$. Set $q_j^A := \frac{r_j^A}{r_j^B}$ and $q_j^B := \frac{r_j^B}{r_j^A}$. Furthermore, given a biased and a de-biased model, let δ_j^A and δ_j^B denote how much higher the de-biased model ranks item j for group A and B compared to the biased model. Then, the *group-specific de-biasing correlation coefficient* for group $G \in \{A, B\}$ measures the Spearman-correlation of q_j^G and δ_j^G over all items, i.e.,

$$\text{DCC}^G := \rho_{\left(q_1^G, \ldots, q_{|J|}^G\right), \left(\delta_1^G, \ldots, \delta_{|J|}^G\right)} \tag{3}$$

where $\rho_{\cdot,\cdot}$ is the Spearman correlation coefficient.

Here, the average difference in ranks δ_j^G represents the mean difference in ranks assigned to an item by the biased model $\text{rank}_{ij}^{\text{biased}}$ and the de-biased model $\text{rank}_{ij}^{\text{de-biased}}$ over all users i in group G:

$$\delta_j^G := \frac{1}{|G|} \sum_{i \in G} \text{rank}_{ij}^{\text{biased}} - \text{rank}_{ij}^{\text{de-biased}} \tag{4}$$

Note that, because the most preferred item has rank 1, the difference is positive if an item is ranked higher by the de-biased model than the biased model. The choice ratios r_j^A and r_j^B should be defined as the ratio of interactions and observations within each group for item j. If such data are unavailable, the choice ratios could be replaced by the total interactions between a group and an item although this measure can be subject to exposure bias [33]. However, the group-specific DCC only captures a one group's perspective. The *de-biasing correlation coefficient* combines the group-specific DCCs into a single measure:

Definition 3.2. Given two group-specific de-biasing correlation coefficients DCC_j^A and DCC_j^B, the *de-biasing correlation coefficient (DCC)* returns their mean as in

$$\text{DCC} := \frac{\text{DCC}^A + \text{DCC}^B}{2}. \tag{5}$$

Assuming that a de-biasing method should reward items inversely to each group's choice ratios, the DCC represents how much of the de-biasing method's effect corresponds to de-biasing. This property is important because the de-biasing should, in addition to making recommendations fairer, retain the optimal ranking computed by the biased model. For example, a "de-biased" model that assigns uniformly random item scores would be unbiased but also undesirable. While such a model would yield a U_{Par} score of zero, and therefore seem like a good candidate, it would yield low accuracy. While such a negative effect would also be identified by accuracy metrics, the DCC enables us to understand whether a drop-off in accuracy is an inherent trade-off of de-biasing or an undesirable side-effect of the specific de-biasing method used. The DCC is bounded between -1 and 1 and a value close to 1 implies a pure de-biasing effect. In this case, we can assume that *all changes to the ranking contribute to de-biasing* in the sense that an item only benefits from the de-biasing more than another item when it *should* benefit more. Positive values closer to zero imply that the de-biasing method has side-effects that do not contribute to de-biasing. Negative scores would imply that the de-biasing method increases bias.

The DCC can also be used in different contexts, not just for gender de-biasing as long as a measure of preference differences, such as choice ratios, is available. The need to specify this measure, however, is also a limitation of the DCC as it requires compressing complex social or cognitive constructs into a single scalar value.

We apply the DCC in Section 3.2 and find that different methods achieve significantly different levels of purity.

3.2 Base Model and De-biasing Method

This section defines the de-biasing methods that we evaluate in section 4. All de-biasing methods assume the same base model architecture, which we introduce first. Then, we list the three de-biasing methods, *Orthogonal bias vector projection*, *Gender vector substraction*, and *Gender deconfounding*.

Base Model architecture. We used the implementation of the multinomial logit-based matrix factorization model by Krause et al. [33]. The model relies on the same basic building blocks as the one from Wang et al. [60] — user embeddings $u_i \in \mathbb{R}^k$, item embeddings

$v_j \in \mathbb{R}^k$ with $k \in \mathbb{N}$ and item intercepts $c_j \in \mathbb{R}$. It computes user-item utilities U_{ij} as follows:

$$U_{ij} \coloneqq u_i \cdot v_j + c_j. \tag{6}$$

However, instead of sampling negatives, it employs a multinomial loss function over the observed user-item interactions that the dataset provides. Hence, our model is conceptually identical to standard matrix factorization models and we can perform the same de-biasing methods as demonstrated below. Additionally, it is also more accurate and, crucially, significantly more robust against the data set's strong exposure bias [33].

Orthogonal bias vector projection [24, 60]:. The method first computes a global bias vector and projects each user embedding orthogonally to it. After training the base model, this method de-biases the user embeddings by subtracting a *global bias vector*. Formally, given user vectors u_i, and user genders

$$g_i \coloneqq \begin{cases} 1 & \text{if user is female} \\ 0 & \text{if user is male} \end{cases}, \tag{7}$$

we first compute the mean female bias vector u^{female} as

$$u^{\text{female}} \coloneqq \frac{\sum_i g_i u_i}{\sum_i g_i} \tag{8}$$

and equivalently a mean male bias vector u^{male}. We then define the global bias vector as

$$u_B \coloneqq \frac{u^{\text{female}} - u^{\text{male}}}{\left\| u^{\text{female}} - u^{\text{male}} \right\|}. \tag{9}$$

Projecting the user vectors u_i on u_B yields the individual de-biased user vectors:

$$u_i^* \coloneqq u_i - (u_i \cdot u_B) u_B. \tag{10}$$

Gender vector substraction:. This method only subtracts the respective gender's mean vector from the individual user embeddings similar to additive operations on word embeddings [42]. After training, we update the user embeddings according to

$$u_i^* = u_i - g_i u^{\text{female}} - (1 - g_i) u^{\text{male}} \tag{11}$$

instead of projecting them as in equation (10).

Gender deconfounding:. While the two former approaches debias the user embeddings, the third attempts to *measure* the effect of gender and to neglect it out during inference, similar to *popularity deconfounding (PD)* [67]. Accordingly, we name this approach *gender deconfounding (GD)*. We introduce learnable gender-item-specific constants c_j^{female} and c_j^{male} that represent the utility that a user perceives from the item due to their gender. Given the users' gender from equation 7, we compute the relevance score μ_{ij} for user i and item j as

$$\mu_{ij} = u_i \cdot v_j + c_j + g_i c_j^{\text{female}} + (1 - g_i) c_j^{\text{male}} \tag{12}$$

After fitting the model, we can compute biased recommendations by ranking preferences according to equation (12) and de-biased recommendations by setting the gender-item-specific constants c_j^{female} and c_j^{male} to zero as in equation 6.

We use one constant per gender to retrieve de-biased recommendations that lie "in the middle" of both gender's preferences. In a regression model, one would usually employ a single variable that equals one if the user has a particular gender, i.e., is female, and that is zero otherwise. In that case, dropping the gender constant during inference would lead to all users receiving stereotypically male recommendations, which would be fair, but undesired in many scenarios. By using two constants and L2-regularizing them, the model assigns the mean popularity of the item to the non-gender specific constant c_j and any gender-specific effects to the gender specific constants c_j^{female} and c_j^{male}. Due to the regularization, we obtain $c_j^{\text{female}} \approx -c_j^{\text{male}}$. Without regularization, the right side of equation 12 is not clearly identified as c_j^{female} and c_j^{male} could absorb some of the effect of c.[2] Upcoming implementations could skip regularization and instead use an equivalent model of shape

$$\mu = u_i \cdot v_j + c_j + (2g_i - 1)\, c_j^{\text{female}}. \tag{13}$$

4 Methods

We examine **RQ1** by replicating the online user study by [60] in which participants rated biased and unbiased college major recommendations. Our survey captures the same and additional constructs as the original study in a 2x2 between-subjects design (Section 4.3). To answer **RQ2**, we present participants with information on gender bias. We address **RQ3** by controlling for gender in our subsequent analysis (Subsection 5). However, our setup addresses necessary adjustments: We perform all assessments on the same dataset with the U_{Par}, $U_{\text{Par}}^{\text{rank}}$, and DCC measures. The biased and de-biased models share the same base architecture (Section 4.2).

4.1 Dataset

We trained our models based on the dataset from [33]. The course catalog contains 100 German course titles inspired by the course offer of a major German educational institution. They cover the categories of Sports, Health, Soft Skills, Media, History, Society, Philosophy, Economy, Science, Personal Development, Psychology, Languages, and Others. Participants made 40 choices from choice sets of four randomly selected courses based on the course's title, resulting in 11,360 choices by 284 attentive participants. 74 participants were male and 210 female. Because the dataset contains strong exposure bias by design, we employed an exposure bias robust model.

Figure 1 shows how often female and male participants chose a course when it was recommended, grouped by category. Female and male users chose inherently differently. Female participants chose from the categories sports that mostly consisted of yoga, Pilates and similar classes, culture, and languages, compared with male participants. Male participants, on the other hand, preferred classes related to media, politics, and soft skills. Soft skills contained relatively many classes that can benefit personal careers, such as leadership classes.

4.2 Model and De-biasing Method

The original study found that the de-biased model performed better than the biased model. However, in the study the biased system recommends *"based on the choices by the people of the same gender"* while the debiased system recommends *"based on the choices by*

[2] In the user study, we employed GD without regularization. In a succeeding offline evaluation, missing regularization did not impact performance in our setup significantly but we advise to employ it regardless.

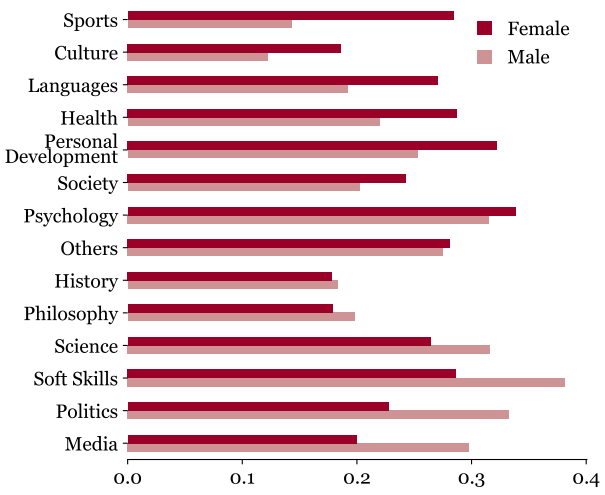


Figure 1: Gender bias in our dataset. The bars represent the mean choice ratios, i.e., the ratios of interactions and observations per category by gender. Female choice ratios are in dark red, male choice ratios in coral.

the people of both genders.", suggesting that the de-biased model learned on, in this case, 67 percent more samples than the biased model, which would artificially boost its accuracy. We evaluated the performance on our dataset and found that the de-biased model performs slightly worse than the biased model. Moreover, we initially found that, in our setup, the approach from Wang et al. [60] does not achieve a significantly non-zero DCC score (Table 1).[3]

Based on our observations, we decided to subsequently assess additional de-biasing methods until one displayed significant de-biasing capability according to DCC. That method would then undergo an additional manual, qualitative assessment. Additionally, we measured the nDCG [58], which is a standard measure for recommendation accuracy, for all methods to understand the potential trade-off between de-biasing and performance. Overall, we evaluated the performance of the following three de-biasing approaches (Section 3):

- **Orthogonal bias vector projection [24, 60]**
- **Gender vector substraction**
- **Gender deconfounding (GD)**

Table 1 contains the averaged results over 100 repetitions. All de-biasing methods performed comparably well in terms of nDCG. The performance loss through de-biasing was around 0.01, which is negligible considering the performance results in the existing study on the same dataset [33]. However, only GD achieved a DCC score close to 1. The method from Wang et al. [60] scored the lowest on DCC, followed by *Gender vector substraction*. GD also scored a lower de-biased U_{Par} and $U_{\text{Par}}^{\text{rank}}$ than the other methods.

[3]An earlier DCC version was used. The DCC from Section 3 returns significantly non-zero scores for all methods. The original study's method still scores the lowest.

After applying a two-sided bootstrapped t-test and multiple test correction, differences between the models were all significant at the $\alpha = 0.001$ level except for nDCG where we observed no significantly different mean values. Note that the bootstrapped t-test does not assume normality [54]. Manual inspection confirmed that GD recommends more stereotypically male courses to women and vice-versa to men. We did not observe any clear de-biasing patterns in the other methods' outputs.

Note that our study's goal was to measure preferences towards or against de-biased recommendations, not to evaluate the best de-biasing method, if such even exists. Hence, we needed to validate our de-biasing strategy, but comparing all possible de-biasing strategies was out of scope. While *GD* out-performed the other approaches, a broader comparison to more de-biasing methods on multiple datasets would be necessary to determine its superiority.
Hyper-parameters. We selected the embedding dimensions and l2 regularization parameters with Bayesian hyper-parameter optimization based on the mango python package [49] and the early stopping criterion from [38]. Any other hyper-parameters were identical to the ones in [33]. We used two thirds of all samples as training data and the rest for validation.

4.3 User study

To understand users' attitudes towards de-biased recommendations, we conducted an online user study with a 2 (*recommender type: biased* vs. *gender de-biased*) x2 (*intervention: yes* vs. *no*) between-subjects design. The study consisted of four core phases: After a brief welcome, data privacy statement, and experiment description, participants underwent (i) interactive *preference elicitation*, optionally an (ii) *intervention*, (iii) viewing *recommendations*, and (iv) a *survey*. We implemented the experiment as an online survey using the open-source python framework oTree [9].
(i) Preference elicitation. The first phase captured the participants' interests to compute corresponding user embedding. The participants subsequently indicated their preference among four randomly drawn courses from the first 50 out of all 100 courses. Each participant made 10 choices in total. This data was then used to train the new user embedding while fixing the existing item embeddings. Because the new user embedding was present in relatively more training samples than the existing users from the dataset were, we had to increase the L2-regularization parameter to achieve a similar embedding length as with the other user-embeddings. A value of 0.3 sufficed.
(ii) Intervention. Next, we presented the intervention group with information on gender bias to investigate **RQ2**. Participants received a brief explanation on how gender bias works and why it may be undesirable in such a recommender system. We kept the explanation concise so that participants would pay attention. Translated from the German version used in the experiment it read:

> *"AI systems learn from historical data and tend to reinforce prevailing patterns of thought in society. For example, they reflect historically evolved gender images. AI-based recommendation systems for further education courses tend to suggest gender-stereotypical learning content. Individual potential may not be able to develop*

Table 1: Performance of the compared de-biasing methods, averaged over 100 repetitions each. Each cell contains the mean value followed by the 25 percent and 75 percent quartiles in parentheses.

De-biasing method	DCC ↑		U_{Par} ↓		$U_{\text{Par}}^{\text{rank}}$ ↓		nDCG ↑	
Biased	–		0.25	(0.21, 0.30)	3.80	(3.23, 4.36)	0.73	(0.72, 0.74)
Orthogonal bias vector projection [24]	0.35	(0.26, 0.45)	0.13	(0.10, 0.16)	2.24	(1.69, 2.75)	0.72	(0.71, 0.73)
Gender vector substraction	0.72	(0.68, 0.77)	0.11	(0.08, 0.13)	1.82	(1.43, 2.15)	0.72	(0.71, 0.73)
GD (Ours)	**0.91**	**(0.90, 0.92)**	**0.07**	**(0.05, 0.08)**	**1.43**	**(1.11, 1.65)**	**0.73**	**(0.72, 0.74)**

> *freely as a result. In the following, we present a recommendation system from which these gender stereotypes have been algorithmically removed."*

The underlined passages were highlighted and contained a tooltip that displayed additional text when hovered, for further explanations. The tooltip for *"AI"* read *"Artificial Intelligence"* and the tooltip for *"gender-stereotypical learning content"* read *"For example, women tend to be recommended stereotypically female and men stereotypically male educational courses."*.

The intervention text stated that the participants are receiving de-biased recommendations. This was only true for half of the participants (*intervention: yes*, *recommender type: gender debiased*), not for the other half (*intervention: yes*, *recommender type: biased*). This way, we could observe the effect of the information on bias on user perception alone.

(iii) Recommendations. After the intervention, we generated the recommendations. Depending on the experiment group, recommendations for participants with the condition *recommender type: biased* got their recommendations from the "standard" recommender, and the ones with *recommender type: gender debiased* from the de-biased recommender. All course recommendations were from the last 50 courses in the dataset. Since the model was trained on the first 50 courses, the model coud not observe any of the participant's preferences for the recommended items and participants did not receive any recommendations for courses they had already indicated their preferences for.

(iv) Survey. Next, we evaluated participants' perceptions of the recommendations and their dependence on predefined control variables. To evaluate the recommendations, we first presented the models' predicted top-6-to-10 and bottom-5 preferences in a randomly shuffled list. We then prompted the participants to re-rank the items based on their preferences. This procedure aimed at assessing whether the model managed to identify the top preferences independent of the mean item quality.

We then asked each user to provide their opinions on their top-5 recommendations to compare our results to the ones from Wang et al. [60] to answer **RQ1** and **RQ2**. For each recommendation, the participants answered questions from the scale developed by Bauer et al. [3] to measure product pleasure. The questions show good internal consistency with a Cronbach's alpha of 0.87 and include the following three questions:

(1) I like the course.
(2) I find the course exciting.
(3) I am interested in the course.

All questions were rated on a seven-point Likert scale ranging from 1: strongly disagree to 7: strongly agree. We slightly adapted the questions to fit our use case and translated them into German. Next, we measured the perceived quality across the top-5 recommended courses using a scale developed by Jones and Pu [27]. The scale shows excellent internal consistency with a Cronbach's alpha of 0.96 and contains the following six questions:

(1) The courses recommended to me were enjoyable.
(2) The courses recommended to me were tailored to my taste.
(3) In general, I am satisfied with the courses recommended to me.
(4) The recommended courses are as good as those I would receive from my friends.
(5) The system's recommendation technology is accurate.
(6) The system understands my tastes and preferences when it comes to education.

To compare our results to the ones of Wang et al. [60], we also used the same questions on the participants' attitude towards gender-stereotypes in career recommendations and their personal inherent gender inequality, as well as their likeliness to use the system again and to recommend the system to other users:

(1) (Q-Stereotype): A gender stereotype in career selection is undesirable since it limits women's and men's capacity to develop their personal abilities.
(2) (Q-DisparityPersonal):
   (a) If I am a female, I do not want to choose a career that is male-dominated (for female participants)
   (b) If I am a male, I do not want to choose a career that is female-dominated (for male participants)
(3) (Q-UseAgain): I would like to use a career recommendation system like this in the future.
(4) (Q-RecommendToOthers): I would like to recommend the system to my friends if it is available.

Lastly, to statistically control the variable of sexism awareness in our analysis, we asked for participants' agreement on a scale about subtle sexism by Swim et al. [53]. With a Cronbachs Alpha of 0.93, the scale shows excellent internal consistency. Items highlighted with '*' are reversed for computation:

(1) *Discrimination against women is no longer a problem in Germany
(2) Women often miss out on good jobs due to sexual discrimination.
(3) *It is rare to see women treated in a sexist manner on television.
(4) *On average, people in our society treat husbands and wives equally.
(5) *Society has reached the point where women and men have equal opportunities for achievement.
(6) It is easy to understand the anger of feminists in Germany.

Table 2: Gender distribution per experiment group.

Recommender	Intervention	Female	Male	**Overall**
Biased	No	97	92	**189**
	Yes	103	87	**190**
Gender de-biased	No	88	102	**190**
	Yes	90	103	**193**
Overall		**378**	**384**	**762**

(7) It is easy to understand why women are still concerned about societal limitations on their opportunities.
(8) *Over the past few years, the government and news media have been showing more concern about the treatment of women than is warranted by women's actual experiences.

Implementation. We implemented the experiment as an online survey using the open-source python framework oTree [9].
Participants. We recruited the participants via Prolific[4]. This allowed us to filter for native German speakers and to achieve a balanced amount of male and female participants. We aimed to collect answers from 800 participants. Prolific assigned the participants randomly into the four experiment groups. The succeeding evaluation only included responses that met the following criteria:

- Participants completed the survey and filled all required fields.
- Participants indicated their gender to be male or female. Because the main focus of this study and the study by Wang et al. [60] is explicitly concerned with male and female gender stereotypes.
- Participants passed the comprehension check. Before indicating their course preferences, participants received a test choice set and had to choose one particular course.
- Participants passed the attention check. Towards the end of the survey, one of the question blocks explicitly asked for the option "I agree".

The experiment was published on April 25th, 2024 at 11:15. The last participant finished on April 28th, 2024 at 13:16. We received 801 initial responses with 762 responses meeting all inclusion criteria. Table 2 shows the numbers of participants with respect to the experiment groups. The mean completion time was 9:55 minutes.

5 Results

5.1 Recommendation Accuracy

Table 3 shows how well the model managed to distinguish between the user's top and bottom preferences. As described in Section 4.3, participants ranked their predicted top 6-10 and bottom five courses. The scores in Table 3 represent how many of the top courses were on average assigned to the predicted group, top 5 or bottom 5. For all groups, the mean scores are above 3.6, well above the expected mean of 2.5 for a random recommendation policy. Hence, both the biased and the unbiased model managed to identify user preferences.

5.2 Participant Ratings

To examine the comparability of our results with the results of Wang et al. [60], we asked our participants to answer the same questions as in the original study (Q-Stereotype, Q-DisparityPersonal, Q-UseAgain, Q-RecommendToOthers).

[4]https://www.prolific.com

Table 3: Recommender performance by experiment group.

Recommender	Intervention	Mean	Standard Deviation
Biased	No	3.80	0.96
	Yes	3.69	0.89
Gender de-biased	No	3.62	1.04
	Yes	3.69	0.87

Figures 2a and 3a show the results from the original paper, Figures 2b, 2c, 3b and 3c those from our replication survey. Results for Q-Stereotype (M=2.71, SD=1.6) and Q-DisparityPersonal (M=2.54, SD=1.54) indicate that participants are rather less biased towards Gender Roles in Career Choices, showing the same trend as Wang et al. [60]. Q-UseAgain (M=4.17, SD=1.64) and Q-RecommendToOthers (M=3.92, SD=1.73) also resemble the trend in Wang et al. [60].

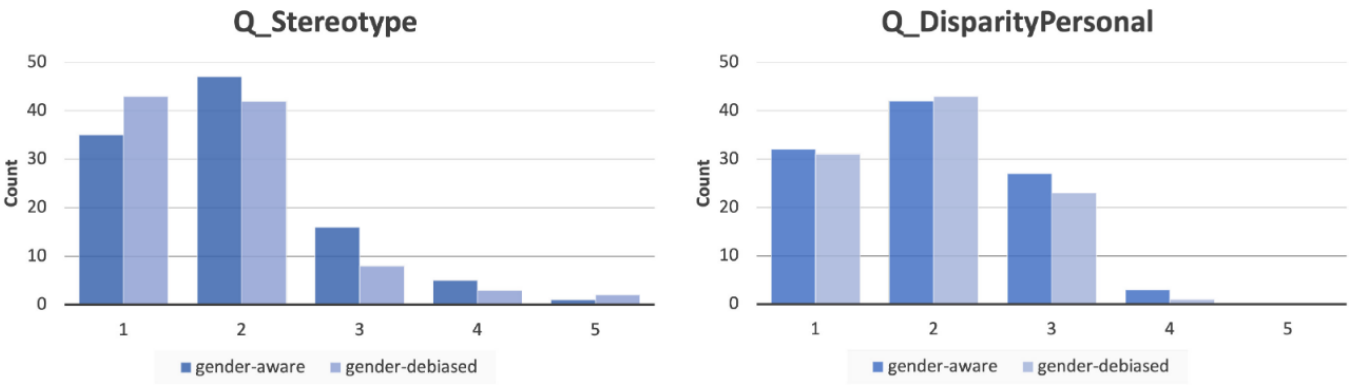
A 2X2X2 ANCOVA was conducted to assess the effects of the independent variables recommender type (*biased*/ *gender de-biased*), gender (*female*/*male*), and whether the participant received an intervention (*yes*/*no*) on the recommender quality. Participants rating on the sexism scale was used as a covariate. The results show that the tendency of participants to rate the quality of the biased recommender higher than the gender de-biased version is a statistically significant effect (F(1,754)=5.16, p<0.05). Males also rated the quality significantly higher than females (F(1,754)=9.29, p<0.01). There is, however, no significant difference in whether participants had an intervention or not (F(1,754)=1.20, p=0.27). Additionally, there is an interaction effect between recommender type and gender (F(1,754)=7.6, p<0.01), with the gender de-biased recommender quality being rated lower by women, than by men (Figure 4a). The detailed results are available in Table 4 in the appendix.

We also conducted a 2X2X2 ANCOVA to assess the effects of the independent variables recommender type (*biased*/*gender de-biased*), gender (*female*/*male*), and whether the participant received an intervention (*yes*/*no*) on the recommended course satisfaction. Participants ratings on the sexism scale was used as a covariate. Just as for the recommender quality, the course satisfaction is also significantly higher for participants using the biased recommender (F(1,754)=6.76, p<0.01). There is no significant difference in whether participants had an intervention or not (F(1,754)=1, p=0.32), or their gender (F(1,754)=2.32, p=0.13). Additionally, there is an interaction effect between recommender type and gender (F(1,754)=9.18, p<0.01). In contrast to the men, the women reported higher satisfaction with the biased recommender than with the gender de-biased recommender (Figure 4b). The detailed results are available in Table 5 in the appendix.

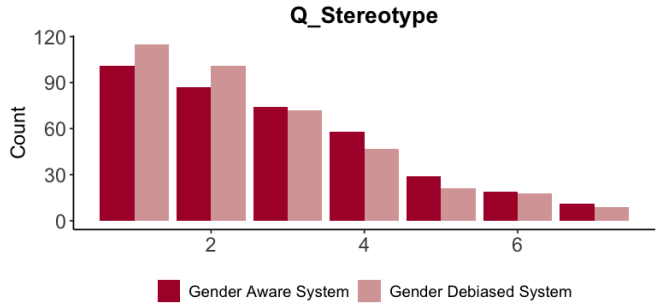
6 Discussion

6.1 Reproducibility of Wang et al. [60]

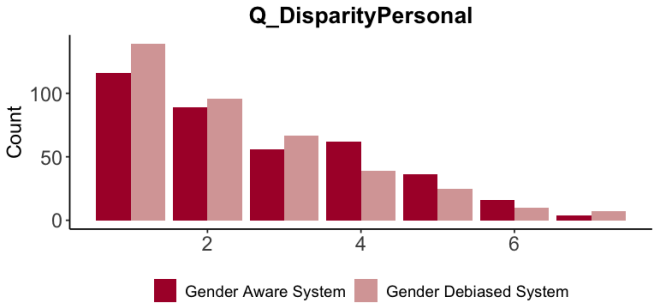
Our main goal was to reproduce the study by Wang et al. [60] with necessary adjustments. However, we find that the results of the user studies are largely consistent. On average, the participants favored the biased recommendations over the de-biased ones. They rated the quality of the biased recommendations significantly higher and the biased model was better able to distinguish between their top



(a) Results from original study by [60]



(b) Q_Stereotype (from 1: Strongly agree or least biased to 7: Strongly disagree or most biased)



(c) Q_DisparityPersonal (from 1: Strongly disagree or least biased to 7: Strongly agree or most biased)

Figure 2: Self-reported beliefs about gender roles in career choices.

and bottom preferences. We also measure similar self-reported believes and effects on recommendation quality as the original study. Regarding the personal beliefs on gender roles in career choice (Section 5.2), we observe a similar distribution of Q-Stereotype and Q-DisparityPersonal. For Q-UseAgain and Q-RecommendToOthers we also observe a similar distribution of participant answers between our and the original study. We observe no significant effect of Q-Stereotype on gender de-biased recommendation acceptance, but one for Q-DisparityPersonal, which is in line with the results from the original study. Our participants' self-assessments show an even less biased distribution. Additionally, we find that more sexism-aware participants exhibit less gender bias, and that women are less biased regarding Q-DisparityPersonal. In both studies and for both questions, the biased recommender has slightly, but not significantly, higher rankings than the gender de-biased version. Overall, our study replicates and confirms the results from [60] and shows that they hold despite different study designs. We answer **RQ1** as follows:

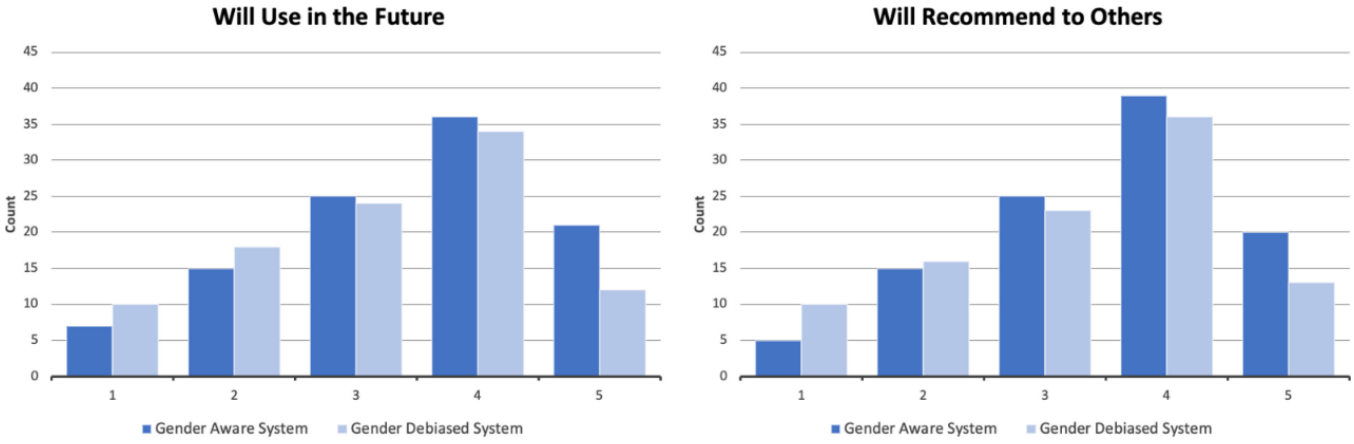
> *We can reproduce the findings of Wang et al. [60] with verified de-biasing methods and comparable models.*

6.2 Effects of gender bias information on user perception

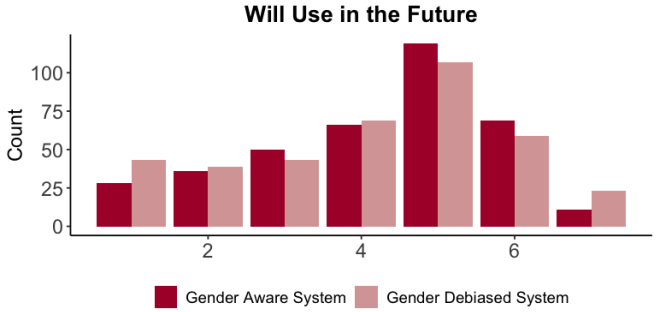
A follow-up questionnaire conducted in [60] indicates that explanations on gender bias, and its implications for recommender systems could help nudge users to accept gender de-biased recommendations better. We presented an intervention, including such an explanation, to every second participant in our online user study (Section 4.3). However, we did not find any significant effect of the explanations on how users perceived recommender quality, their satisfaction with the course recommendations, or how fitting they find their recommendations to their interests. Yet, although not significantly, users having received an intervention showed equal or higher satisfaction with their recommendations throughout all groups and concepts in our study. We therefore answer **RQ2** with:

> *We did not find any significant evidence that concise and high-level explanations on gender bias influence user perception of gender de-biased educational recommendations.*

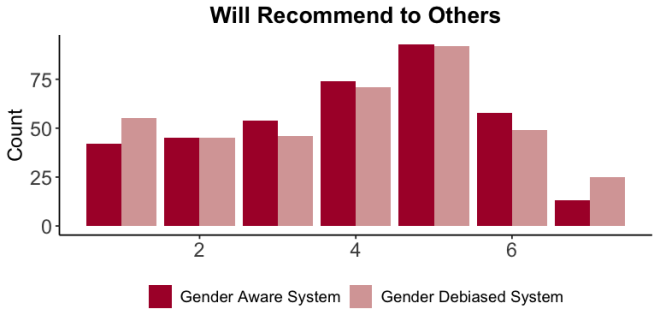
Several perspectives could explain our observation. On the one hand, a different prompt design could have achieved different results. [56] showed that exposing people to scientific evidence about stereotypes to reduce their biases can actually cause the opposite effect, due to these cues' high social complexity. A successful intervention could therefore require a different design. We chose a short and high-level prompt to ensure that participants pay attention. However, our prompt could have benefited from more length, more detail, or simpler language. On the other hand, explicit education on gender bias could not be the best way of improving user perception. The hypothesis that such information improves acceptance of de-biased recommendations relied on the responses



(a) Results from original study by [60]



(b) Q_UseAgain (from 1: Strongly disagree or least satisfied to 7: Strongly agree or most satisfied)



(c) Q_RecommendToOthers (from 1: Strongly disagree or least satisfied to 7: Strongly agree or most satisfied)

Figure 3: Participant usability ratings

to the questionnaire by [60], not on established theory. Moreover, the survey sample size of 20 participants was relatively small. Other participants could have suggested different design changes.

6.3 Gender differences

In extension to Wang et al. [60], we investigated whether male and female participants perceived gender de-biased recommendations differently compared to biased recommendations. Indeed, male participants did not indicate any preference for biased recommendations. Instead, they slightly, although insignificantly, preferred the de-biased recommendations (Figure 4). We therefore answer **RQ3** with:

> *Only female users significantly preferred biased recommendations. Male users slightly, but insignificantly preferred de-biased recommendations.*

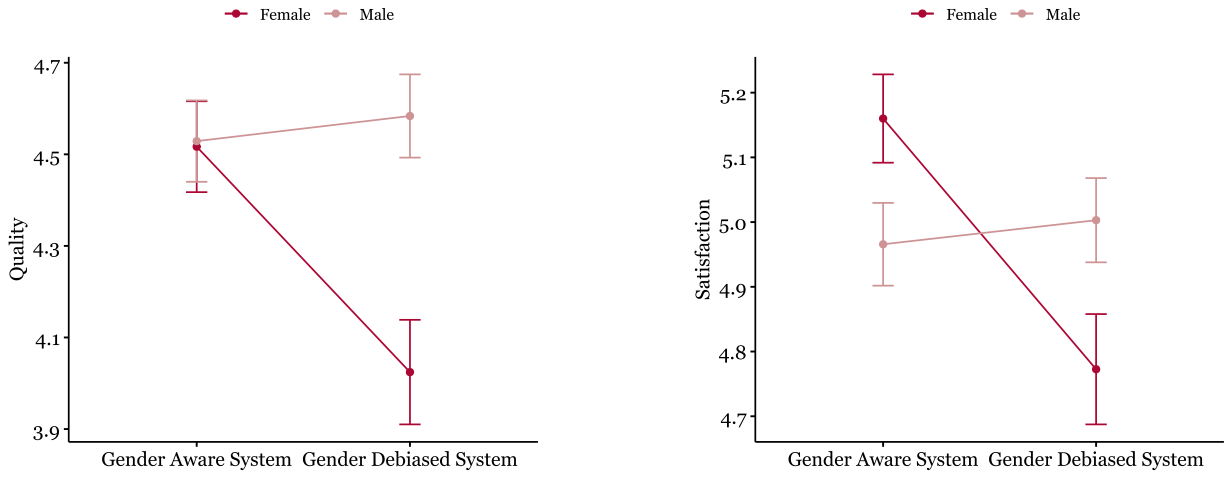
The observation appears to contradict the self-reported beliefs about gender roles in career choices by the participants. In Q-DisparityPersonal, women's answers indicate that they are more willing to work in a male-dominated career than men in a female-dominated career. The contradiction could be due to the fact that the educational recommendations were not career-oriented.

Alternatively, the group sizes in the training data could explain our observations. Females made up the majority of the training data set at 74% of all participants. Hence, the model could predict based on twice as many similar preferences for "typically" female choices behavior than typically male choices. Moreover, the female training samples could have outweighed the male samples during training. Hence, female users could prefer the biased model because they are the *majority*. Similar model behavior has been observed in-between different groups in the context of exposure bias [39].

A more speculative explanation for women gravitating toward stereotypically female educational recommendations could be that women tend to be less confident than men when envisioning themselves in a counter-stereotypical environment. For example, men are more likely to apply for a job even if they do not meet all the qualifications, whereas women typically only apply when they feel particularly qualified [55]. Women may be more comfortable in familiar environments where they possess all required competences while men may be more open towards areas that they are unexperienced in.

6.4 Limitations

Our study was limited to German participants. It only measured gender stereotypes of parts of the German society. With [60] obtaining similar results from their US study and the cultural and political similarities, we assume that these tendencies also hold true in at least Nothern America and Western Europe. On a global scale,



(a) Recommender quality by experiment group.

(b) Course satisfaction by experiment group.

Figure 4: Participants' ratings on recommendation quality and satisfaction with their top-5 recommendations.

gender-based division of labor is consistent throughout all cultures, but to very different extents, so that the observed effects may differ more strongly around the globe [21].

Moreover, our results, just as those of the replicated study, rely on stated preferences. Revealed preferences could differ. Users could decide differently in real-world scenarios where their educational choices affect their identities and how others perceive them, especially considering the investment of taking a time-intensive course or even an expensive major. In our lab environment, the benefit of appearing unbiased by selecting non-stereotypical courses could outweigh the choices' imaginary implications.

Lastly, we only presented one concise and high-level explanation on gender bias to the intervention group. A different design could have better introduced participants to the complex topic that gender bias is. Future studies could also include a comprehension task and control for the participants' understanding.

6.5 Implications

Our key observation, that users on average dislike gender de-biased recommendations, is in-line with previous findings on interactions with search engines from [30] with the exception that in our study, participant's gender significantly influenced their perception of recommendation quality. Hence, men and women could have similar biases about the outside world that affect their own preferences differently. Alternatively, group sizes in the training data set could have played a significant role so that de-biasing affected the majority and minority group differently. Last, the fact that the biased system is optimized for acceptance could imply that de-biasing methods always decrease user satisfaction.

Much research exists on measuring and mitigating gender bias in recommendations. We demonstrate that removing the gender bias from *recommendations* is not enough because de-biasing can decrease user satisfaction, even when users are aware of the bias. We attempted to remove the gender bias from participants through short explanations on what gender bias is and why it can be harmful. However, despite participants from a previous study stating that such explanations would help them overcome gender bias, we did not find any significant effect on satisfaction. One approach to reduce people's gender biases is embodiment [25], and perhaps in some settings, similar approaches could be deployed. For example, when recommending career options, users could view examples of personas that they identify with in contra-stereotypical occupations. Without a solution, commercial recommendations providers could opt-out of de-biasing their recommendations. This could, in turn, fuel the users' biases further. Public, not-for-profit recommendation providers that want to de-bias their recommendations require efficient methods for retaining user acceptance.

Our findings also revealed that de-biasing affects genders benefit differently, perhaps based on relative group sizes. The male minority responded more favorable of the de-biased recommendations than the female majority. We do not know whether gender or relative group sizes are responsible. However, future research on gender de-biasing and potentially on general de-biasing, should group their results by gender or groups in general. Otherwise, majority groups could skew the results in their favor.

We showed that existing de-biasing measures do not always paint the whole picture. Our measure, the DCC, can help researchers and practitioners better assess de-biasing methods. Our proposed de-biasing method GD out-performed the other methods in our setup and could potentially proof superior to existing approaches in a broader comparison.

While not included in table 1, we also evaluated GD with active gender-specific intercept terms during development, effectively providing the base model with the true gender features. We found that this approach *increases* gender bias significantly over the base model. Hence, controlling for sensitive features could help the model to learn stereotypical relationships and bias them even stronger.

6.6 Future Research Directions

Our results open towards several important research questions. First, in our study, female users preferred the biased recommendations and they were in the majority group. We do not know whether their gender or them being the majority group drives this preference. Further research could employ a gender-balanced dataset, or train one model on predominantly female data and one on predominantly male data and then compare the two to investigate whether women generally prefer biased recommendations. Second, our study and the original study [60] ask participants to state their preferences. Field studies are necessary to verify that this effect translates to real-world environments. Last, the presented participants with brief, high-level information on gender bias. If and which other designs, for example, different content length and complexity, video or image content, or even entirely different concepts, could yield a stronger effect on users remains a challenge for future research.

7 Conclusion

This study replicates the work of Wang et al. [60] on whether users perceive gender de-biased recommendations better than gender-biased recommendations while addressing important limitations. Additionally, it investigates whether providing users with information on gender bias changes how they perceive gender de-biased recommendations. We perform an online user study with 800 participants in a 2x2 between-subjects design. In contrast to the original study, we rigorously assess the de-biasing method, find that it yields unsatisfying results in our setting and identify better performing methods. We propose a novel de-biasing measure, the DCC, to evaluate to which extend an assumed de-biasing-method actually contributes to fairness. We find that a Popularity deconfounding-inspired method, GD, improves demographic parity significantly better than the method from the original study. Overall, our findings are consistent with the ones from Wang et al. [60], i.e. on average the participants prefer biased recommendations. However, we observe that only female participants significantly prefer biased recommendations. Male participants slightly prefer de-biased recommendations, although insignificantly. Hence, preferences differ by gender and the majority female group skews the overall result. Future research should investigate whether gender, majority group-status, or both drive this preference. We also do not find significant evidence that explanations on gender bias improve users' perception of de-biased recommendations. Differently designed explanations or different approaches could have greater effects on user behavior.

Acknowledgments

This contribution originates from the research project KUPPEL funded by the Federal Ministry of Education and Research (BMBF), ref. no. 21INVI0803.

References

[1] Oscar Alvarado, Nyi Nyi Htun, Yucheng Jin, and Katrien Verbert. 2022. A systematic review of interaction design strategies for group recommendation systems. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–51.

[2] Ashwathy Ashokan and Christian Haas. 2021. Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management* 58, 5 (2021), 102646.

[3] Hans H Bauer, Nicola E Sauer, and Christine Becker. 2006. Investigating the relationship between product involvement and consumer decision-making styles. *Journal of Consumer Behaviour* 5, 4 (2006), 342–354.

[4] Joeran Beel and Haley Dixon. 2021. The 'Unreasonable' Effectiveness of Graphical User Interfaces for Recommender Systems. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 22–28. https://doi.org/10.1145/3450614.3461682

[5] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery, New York, NY, USA, 2212–2220. https://doi.org/10.1145/3292500.3330745

[6] Dirk Bollen, Bart P. Knijnenburg, Martijn C. Willemsen, and Mark Graus. 2010. Understanding choice overload in recommender systems. In *Proceedings of the Fourth ACM Conference on Recommender Systems* (Barcelona, Spain) *(RecSys '10)*. Association for Computing Machinery, New York, NY, USA, 63–70. https://doi.org/10.1145/1864708.1864724

[7] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*. 13–18. https://doi.org/10.1109/ICDMW.2009.83

[8] André Calero Valdez, Martina Ziefle, and Katrien Verbert. 2016. HCI for Recommender Systems: the Past, the Present and the Future. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16)*. Association for Computing Machinery, New York, NY, USA, 123–126. https://doi.org/10.1145/2959100.2959158

[9] Daniel L. Chen, Martin Schonger, and Chris Wickens. 2016. oTree—An open-source platform for laboratory, online, and field experiments. 9 (2016), 88–97. https://doi.org/10.1016/j.jbef.2015.12.001

[10] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems* 41, 3 (2023), 1–39.

[11] Zhilong Chen, Jinghua Piao, Xiaochong Lan, Hancheng Cao, Chen Gao, Zhicong Lu, and Yong Li. 2022. Practitioners versus users: A value-sensitive evaluation of current industrial recommender system design. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–32.

[12] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (Boston, Massachusetts, USA) *(RecSys '16)*. Association for Computing Machinery, New York, NY, USA, 191–198. https://doi.org/10.1145/2959100.2959190

[13] Jamell Dacon and Haochen Liu. 2021. Does Gender Matter in the News? Detecting and Examining Gender Bias in News Articles. In *Companion Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 385–392. https://doi.org/10.1145/3442442.3452325

[14] Jeffrey Dastin. 2022. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*. Auerbach Publications, 296–299.

[15] Karlijn Dinnissen and Christine Bauer. 2022. Fairness in Music Recommender Systems: A Stakeholder-Centered Mini Review. *Frontiers in Big Data* 5 (2022). https://doi.org/10.3389/fdata.2022.913608

[16] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. 2022. Online certification of preference-based fairness for personalized recommender systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 6532–6540.

[17] Alice H Eagly and Wendy Wood. 2012. Social role theory. *Handbook of theories of social psychology* 2 (2012), 458–476.

[18] Daniel Fallman. 2003. Design-oriented human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 225–232. https://doi.org/10.1145/642611.642652

[19] Andres Ferraro, Michael D. Ekstrand, and Christine Bauer. 2024. It's Not You, It's Me: The Impact of Choice Models and Ranking Strategies on Gender Imbalance in Music Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems* (Bari, Italy) *(RecSys '24)*. Association for Computing Machinery, New York, NY, USA, 884–889. https://doi.org/10.1145/3640457.3688163

[20] Andres Ferraro, Xavier Serra, and Christine Bauer. 2021. Break the Loop: Gender Imbalance in Music Recommenders. In *Proceedings of the 2021 Conference on*

Human Information Interaction and Retrieval (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 249–254. https://doi.org/10.1145/3406522.3446033

[21] Peter Glick and Susan T Fiske. 1999. Gender, power dynamics, and social interaction. *Revisioning gender* 5 (1999), 365–398.

[22] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[23] Blake Huebner, Thomas Elmar Kolb, and Julia Neidhardt. 2024. Evaluating Group Fairness in News Recommendations: A Comparative Study of Algorithms and Metrics. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, 337–346. https://doi.org/10.1145/3631700.3664897

[24] Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James Foulds. 2021. Debiasing Career Recommendations with Neural Fair Collaborative Filtering. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 3779–3790. https://doi.org/10.1145/3442381.3449904

[25] Marie Jarrell, Reza Ghaiumy Anaraky, Bart Knijnenburg, and Erin Ash. 2021. Using Intersectional Representation & Embodied Identification in Standard Video Game Play to Reduce Societal Biases. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 446, 18 pages. https://doi.org/10.1145/3411764.3445161

[26] Ray Jiang, Silvia Chiappa, Tor Lattimore, András György, and Pushmeet Kohli. 2019. Degenerate Feedback Loops in Recommender Systems. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 383–390. https://doi.org/10.1145/3306618.3314288

[27] Nicolas Jones and Pearl Pu. 2008. User acceptance issues in music recommender systems. (2008).

[28] Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems* 33, 1 (2012), 1–33.

[29] Chen Karako and Putra Manggala. 2018. Using Image Fairness Representations in Diversity-Based Re-ranking for Recommendations. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization* (Singapore, Singapore) *(UMAP '18)*. Association for Computing Machinery, New York, NY, USA, 23–28. https://doi.org/10.1145/3213586.3226206

[30] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520

[31] Mesut Kaya, Derek Bridge, and Nava Tintarev. 2020. Ensuring Fairness in Group Recommendations by Rank-Sensitive Balancing of Relevance. In *Proceedings of the 14th ACM Conference on Recommender Systems* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 101–110. https://doi.org/10.1145/3383313.3412232

[32] Simone Kopeinik, Martina Mara, Linda Ratz, Klara Krieg, Markus Schedl, and Navid Rekabsaz. 2023. Show me a "Male Nurse"! How Gender Bias is Reflected in the Query Formulation of Search Engine Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 137, 15 pages. https://doi.org/10.1145/3544548.3580863

[33] Thorsten Krause, Alina Deriyeva, Jan H. Beinke, Gerrit Y. Bartels, and Oliver Thomas. 2024. Mitigating Exposure Bias in Recommender Systems—A Comparative Analysis of Discrete Choice Models. *ACM Trans. Recomm. Syst.* 3, 2, Article 19 (Nov. 2024), 37 pages. https://doi.org/10.1145/3641291

[34] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. 2023. Fairness in Recommendation: Foundations, Methods, and Applications. *ACM Trans. Intell. Syst. Technol.* 14, 5, Article 95 (Oct. 2023), 48 pages. https://doi.org/10.1145/3610302

[35] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards Personalized Fairness based on Causal Notion. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1054–1063. https://doi.org/10.1145/3404835.3462966

[36] Haifeng Liu, Yukai Wang, Hongfei Lin, Bo Xu, and Nan Zhao. 2022. Mitigating sensitive data exposure with adversarial learning for fairness recommendation systems. *Neural Computing and Applications* 34, 20 (2022), 18097–18111. https://doi.org/10.1007/s00521-022-07373-4

[37] Chieh Lo, Hongliang Yu, Xin Yin, Krutika Shetty, Changchen He, Kathy Hu, Justin M Platz, Adam Ilardi, and Sriganesh Madhvanath. 2021. Page-level Optimization of e-Commerce Item Recommendations. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, Netherlands) *(RecSys '21)*. Association for Computing Machinery, New York, NY, USA, 495–504. https://doi.org/10.1145/3460231.3474242

[38] Anastasia Makarova, Huibin Shen, Valerio Perrone, Aaron Klein, Jean Baptiste Faddoul, Andreas Krause, Matthias Seeger, and Cedric Archambeau. 2021. Overfitting in Bayesian Optimization: an empirical study and early-stopping solution. In *2nd Workshop on Neural Architecture Search (NAS 2021 collocated with the 9th ICLR 2021)* (2021-04). https://doi.org/10.3929/ethz-b-000521574 Accepted: 2022-01-04T08:49:49Z.

[39] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback Loop and Bias Amplification in Recommender Systems. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 2145–2148. https://doi.org/10.1145/3340531.3412152

[40] Masoud Mansoury, Himan Abdollahpouri, Jessie Smith, Arman Dehpanah, Mykola Pechenizkiy, and Bamshad Mobasher. 2020. Investigating potential factors associated with gender discrimination in collaborative recommender systems. In *The thirty-third international flairs conference*.

[41] Alessandro B. Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management* 58, 5 (2021), 102666. https://doi.org/10.1016/j.ipm.2021.102666

[42] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*. 746–751.

[43] Jahna Otterbacher, Jo Bates, and Paul Clough. 2017. Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6620–6631. https://doi.org/10.1145/3025453.3025727

[44] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2022. Fairness in rankings and recommendations: an overview. *The VLDB Journal* 31, 3 (2022), 431–458. https://doi.org/10.1007/s00778-021-00697-y

[45] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. 2019. 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* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 231–239. https://doi.org/10.1145/3289600.3291002

[46] Katherine-Marie Robinson, Violet Turri, Carol J Smith, and Shannon K Gallagher. 2024. Tales from the Wild West: Crafting Scenarios to Audit Bias in LLMs. In *CHI Conference on Human Factors in Computing Systems*.

[47] Clara Rus, Jeffrey Luppes, Harrie Oosterhuis, and Gido H. Schoenmacker. 2022. Closing the Gender Wage Gap: Adversarial Fairness in Job Recommendation. https://doi.org/10.48550/arXiv.2209.09592 arXiv:2209.09592 [cs]

[48] Joni Salminen, Chang Liu, Wenjing Pian, Jianxing Chi, Essi Häyhänen, and Bernard J Jansen. 2024. Deus Ex Machina and Personas from Large Language Models: Investigating the Composition of AI-Generated Persona Descriptions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 510, 20 pages. https://doi.org/10.1145/3613904.3642036

[49] Sandeep Singh Sandha, Mohit Aggarwal, Igor Fedorov, and Mani Srivastava. 2020. Mango: A python library for parallel hyperparameter tuning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3987–3991.

[50] Shrikant Saxena and Shweta Jain. 2024. Exploring and mitigating gender bias in book recommender systems with explicit feedback. *Journal of Intelligent Information Systems* (2024), 1–22.

[51] Katie Seaborn, Shruti Chandra, and Thibault Fabre. 2023. Transcending the "Male Code": Implicit Masculine Biases in NLP Contexts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 138, 19 pages. https://doi.org/10.1145/3544548.3581017

[52] Yu Shi, Myunghwan Kim, Shaunak Chatterjee, Mitul Tiwari, Souvik Ghosh, and Rómer Rosales. 2016. Dynamics of Large Multi-View Social Networks: Synergy, Cannibalization and Cross-View Interplay. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1855–1864. https://doi.org/10.1145/2939672.2939814

[53] Janet K. Swim, Kathryn J. Aikin, Wayne S. Hall, and Barbara A. Hunter. 1995. Sexism and racism: Old-fashioned and modern prejudices. *Journal of Personality and Social Psychology* 68, 2 (1995), 199–214. https://doi.org/10.1037/0022-3514.68.2.199 Place: US Publisher: American Psychological Association.

[54] Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability* 57, 1 (1993), 1–436.

[55] Deanne Tockey and Maria Ignatova. 2019. Gender Insights Report: How women find jobs differently. *LinkedIn Talent Solutions, https://business. linkedin. com/content/dam/me/business/en-us/talent-solutions-lodestone/body/pdf/Gender-Insights-Report. pdf* (2019).

[56] Carlo Tomasetto and Sara Appoloni. 2013. A lesson not to be learned? Understanding stereotype threat does not protect women from stereotype threat. *Social Psychology of Education* 16, 2 (2013), 199–213. https://doi.org/10.1007/s11218-012-9210-6

[57] A Tuckett and F Aldridge. 2009. The NIACE Survey on Adult Participation in Learning 2009: Narrowing Participation.

[58] Hamed Valizadegan, Rong Jin, Ruofei Zhang, and Jianchang Mao. 2009. Learning to Rank by Optimizing NDCG Measure. In *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), Vol. 22. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2009/file/b3967a0e938dc2a6340e258630febd5a-Paper.pdf

[59] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-generated Reference Letters. *arXiv preprint arXiv:2310.09219* (2023).

[60] Clarice Wang, Kathryn Wang, Andrew Y. Bian, Rashidul Islam, Kamrun Naher Keya, James Foulds, and Shimei Pan. 2023. When Biased Humans Meet Debiased AI: A Case Study in College Major Recommendation. *ACM Trans. Interact. Intell. Syst.* 13, 3, Article 17 (Sept. 2023), 28 pages. https://doi.org/10.1145/3611313

[61] Jane Webster and Richard T Watson. 2002. Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly* (2002), xiii–xxiii.

[62] Candace West and Don H Zimmerman. 1987. Doing gender. *Gender & society* 1, 2 (1987), 125–151.

[63] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairness-aware News Recommendation with Decomposed Adversarial Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 5 (May 2021), 4462–4469. https://doi.org/10.1609/aaai.v35i5.16573

[64] Yao Wu, Jian Cao, and Guandong Xu. 2023. Fairness in recommender systems: evaluation approaches and assurance strategies. *ACM Transactions on Knowledge Discovery from Data* 18, 1 (2023), 1–37.

[65] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. 30 (2017). https://proceedings.neurips.cc/paper_files/paper/2017/file/e6384711491713d29bc63fc5eeb5ba4f-Paper.pdf

[66] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research* 20, 75 (2019), 1–42.

[67] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/3404835.3462875

A ANCOVA regression tables

Table 4: ANCOVA regression table for the effects on the recommender quality.

	Sum Sq	Df	F value	Pr(>F)
debiased	9.53	1	5.16	0.0234
intervention	2.22	1	1.20	0.2728
sexismAwareness	2.37	1	1.28	0.2576
survey.1.player.gender	17.15	1	9.29	0.0024
debiased:intervention	0.35	1	0.19	0.6615
debiased:survey.1.player.gender	14.03	1	7.60	0.0060
intervention:survey.1.player.gender	2.81	1	1.52	0.2176
Residuals	1392.43	754		

Table 5: ANCOVA regression table for the effects on the recommended course satisfaction.

	Sum Sq	Df	F value	Pr(>F)
debiased	6.40	1	6.76	0.0095
intervention	0.94	1	1.00	0.3187
sexismAwareness	6.33	1	6.70	0.0099
survey.1.player.gender	2.19	1	2.32	0.1284
debiased:intervention	0.99	1	1.04	0.3074
debiased:survey.1.player.gender	8.68	1	9.18	0.0025
intervention:survey.1.player.gender	0.23	1	0.24	0.6230
Residuals	712.93	754		