NÖHÜ Müh. Bilim. Derg. / NOHU J. Eng. Sci., Erken Görünüm / InPress
Niğde Ömer Halisdemir Üniversitesi Mühendislik Bilimleri Dergisi
Niğde Ömer Halisdemir University Journal of Engineering Sciences
Araştırma makalesi / Research article
www.dergipark.org.tr/tr/pub/ngumuh / www.dergipark.org.tr/en/pub/ngumuh

Gender-aware fairness in iterative recommender systems: A simulation study on popularity bias

İteratif öneri sistemlerinde cinsiyete-duyarlı adalet: Popülerlik yanlılığı üzerine bir simülasyon çalışması

Yildiz Zoralioglu[1], Emre Yalcin[2,*]

[1] *Sivas Cumhuriyet University, Graduate School of Natural and Applied Sciences, 58010, Sivas, Türkiye*
[2] *Sivas Cumhuriyet University, Computer Engineering Department, 58010, Sivas, Türkiye*

Abstract

This study examines how gender-based disparities emerge in recommender systems through feedback loops. While fairness has been studied in static settings, little is known about how repeated user-system interactions impact different demographic groups over time. To address this, we utilize a dynamic simulation framework using synthetic interactions and ten feedback iterations. Based on the MovieLens-1M dataset, users are grouped by gender and evaluated using metrics such as calibration, diversity, and long-tail exposure. Results show that female users consistently receive less favorable outcomes, with popularity bias measures (GAP, MRMC) indicating a growing disadvantage over time. Diversity and novelty scores also decline more sharply for women. These findings suggest that feedback loops can reinforce existing inequalities in recommender systems. The employed framework provides a valuable tool for analyzing the evolution of fairness across iterations and highlights the need for gender-sensitive algorithms that maintain fairness over time.

Keywords: Gender fairness, Recommender systems, Feedback loop, Popularity bias, Demographic disparity.

Öz

Bu çalışma, öneri sistemlerinde geri besleme döngüleri yoluyla cinsiyete dayalı eşitsizliklerin nasıl ortaya çıktığını incelemektedir. Adalet konusu durağan ortamlarda araştırılmış olsa da yinelenen kullanıcı-sistem etkileşimlerinin zaman içinde farklı demografik grupları nasıl etkilediği hakkında çok az bilgi bulunmaktadır. Bu durumu ele almak için, sentetik etkileşimler ve on geri besleme iterasyonu içeren dinamik bir simülasyon çerçevesi kullanılmıştır. MovieLens-1M veri kümesine dayalı olarak kullanıcılar cinsiyete göre gruplanmış ve kalibrasyon, çeşitlilik ve uzun kuyruk içeriklere erişim gibi metriklerle değerlendirilmiştir. Sonuçlar, kadın kullanıcıların sistemden sürekli olarak daha olumsuz sonuçlar aldığını göstermekte; GAP ve MRMC gibi popülerlik yanlılığı metrikleri ise zamanla artan bir dezavantajı ortaya koymaktadır. Ayrıca, kadın kullanıcılar için çeşitlilik ve yenilik skorlarının daha keskin bir şekilde düştüğü gözlemlenmiştir. Bu bulgular, geri besleme döngülerinin öneri sistemlerinde mevcut eşitsizlikleri pekiştirebileceğini ortaya koymakta ve zaman içinde adaleti koruyacak cinsiyete duyarlı algoritmalara duyulan ihtiyacı vurgulamaktadır.

Anahtar kelimeler: Cinsiyet adaleti, Öneri sistemleri, Geri besleme döngüsü, Popülerlik yanlılığı, Demografik eşitsizlik.

1 Introduction

Recommender systems (RSs) have become pivotal in addressing the issue of information overload by delivering personalized content tailored to individual users [1]. These systems are widely applied across e-commerce, digital media, education, and healthcare sectors, improving user experiences through context-aware suggestions [2, 3]. Major platforms like Netflix, Spotify, and Amazon utilize RSs to analyze user behavior, providing personalized recommendations for movies, music, and products, thereby enhancing user engagement.

Collaborative filtering (CF) has emerged as one of the dominant techniques within RSs. These methods can be broadly categorized into two types: user-based CF, which predicts preferences by identifying patterns in user behavior, and item-based CF, which relies on the similarity between items. While these methods have proven effective, they face a significant challenge: popularity bias. This bias occurs when frequently interacted items are recommended over less popular ones, resulting in a skewed distribution of recommendations and reduced diversity [4, 5]. As popular items dominate, less-known items often fail to be represented, further exacerbating the problem [6].

Popularity bias not only diminishes the variety of recommended content but also hampers the discovery of new or niche items. This results in a lack of diversity in the content presented to users, negatively impacting their overall satisfaction with the system [7]. Moreover, this imbalance affects both content creators, whose niche products struggle

* Sorumlu yazar / Corresponding author, e-posta / e-mail: eyalcin@cumhuriyet.edu.tr
Geliş / Recieved: 28.03.2025 Kabul / Accepted: 27.06.2025 Yayımlanma / Published: xx.xx.20xx
doi: 10.28948/ngumuh.1667487

for visibility, and service providers, who face declining user engagement and satisfaction. The focus on popular items leads to a narrowing of content exposure, where recommendations tend to be repetitive, and the variety of items shrinks [8].

Another critical issue arising from popularity bias is the calibration problem, where RSs fail to match users' true preferences accurately. As popular items dominate recommendations, the system's outputs may become miscalibrated, leading to less personalized suggestions that don't reflect users' true interests [9]. This misalignment between recommended items and user preferences underscores the need for more accurate calibration within RSs to enhance the relevance and fairness of suggestions.

Recent research has increasingly focused on how popularity bias differentially impacts user groups, particularly through demographic attributes such as gender. While prior studies have identified distinct engagement patterns with popular content between male and female users, they have predominantly relied on static evaluations, neglecting the dynamic nature of real-world recommender systems [10]. Specifically, how feedback loops amplify or mitigate gender-based disparities in recommendation outcomes over time remains largely unexplored. This gap is critical, as iterative interactions between users and systems can compound biases and lead to systemic unfairness. Addressing this limitation, our study contributes a dynamic, gender-aware evaluation framework that captures the evolving nature of fairness in recommender systems, highlighting structural disadvantages faced by female users as feedback loops progress.

In this study, we employ a framework that examines the impact of popularity bias on male and female users across multiple recommendation cycles. Our main contributions include:

- We utilize a dynamic simulation framework with synthetic feedback to monitor the evolution of recommendation fairness over time.
- We conduct the first gender-specific longitudinal analysis of HPF, MMMF, and VAECF, showing how feedback loops affect fairness, calibration, and accuracy.
- We analyze proportional changes across iterations to reveal gender-specific learning patterns and inform the design of fairness-aware systems.

The remainder of this paper is structured as follows: Section 2 reviews related works on popularity bias, fairness, and feedback mechanisms in RSs. Section 3 introduces the proposed simulation environment methodology for modeling iterative feedback loops and outlines the experimental setup, including the datasets, CF algorithms, and evaluation metrics used. Section 4 presents the results and discusses the key insights gained. Section 5 discusses the limitations of the utilized framework. Finally, Section 6 concludes the paper, summarizing the findings and suggesting directions for future research.

2 Related work

Popularity bias in RSs is well-documented, as algorithms disproportionately favor popular items while often ignoring less-known or niche alternatives [11-13]. This imbalance restricts users' access to diverse content, reduces user satisfaction, and creates systemic disadvantages for content providers [14, 15].

Researchers have proposed three major strategies to mitigate popularity bias: pre-processing, in-processing, and post-processing approaches. Pre-processing methods modify the user–item interaction matrix to reduce inherent biases and have also been applied to incorporate privacy-aware or fairness-sensitive filtering mechanisms [16]. For instance, privacy-preserving collaborative recommenders can mitigate popularity bias by transforming user-item interactions before model training. Recent studies indicate that such pre-processing techniques effectively limit bias amplification while maintaining recommendation performance [17]. In-processing alters algorithmic learning processes to improve fairness [18]. Post-processing techniques re-rank recommendations to demote overrepresented items or promote underrepresented ones [19-21].

In addition to traditional accuracy metrics, recent research have emphasized the importance of beyond-accuracy measures such as fairness, diversity, novelty, and coverage [22, 23]. Abdollahpouri [24] categorize fairness into three dimensions: consumer fairness (C-fairness), provider fairness (P-fairness), and stakeholder fairness (S-fairness).

Studies have shown that not all users are affected equally by popularity bias. Individual user characteristics, such as personality traits, interaction levels, or profile consistency, significantly influence the level of bias experienced [11, 25]. For instance, less extroverted or novelty-averse users are likelier to receive less accurate and less diverse recommendations.

A growing body of work investigates how demographic attributes, such as gender, age, and cultural background, impact the quality and fairness of recommendations. Ekstrand et al. [16] revealed significant differences in rating behavior between male and female users in book recommendation platforms. Ferwerda et al. [18] observed that female users prefer more diverse music content, yet recommender systems often fail to reflect this preference. Deldjoo et al. [26] found that recommendation quality varies by gender, contributing to systemic inequity. To address this, group-sensitive approaches, such as group-based calibration, have been proposed, which improve fairness and satisfaction [27].

Another critical area in recent research involves feedback loops. These loops occur when user interactions influence future recommendations, often reinforcing popularity bias over time. Chaney et al. [28] and Kowald et al. [15] demonstrated that feedback loops can shrink the recommendation space and diminish diversity. Mansoury et al. [29] showed that continuous user interaction with popular items amplifies systemic bias. Krauth et al. [30] used dynamic simulation models to analyze how such loops

evolve, revealing that these effects disproportionately impact different user groups.

Despite these advances, relatively few studies have examined how demographic subgroups, particularly based on gender, are dynamically influenced by feedback loops in recommender systems. Our work addresses this gap by categorizing users by gender and systematically evaluating how each group is impacted in terms of accuracy, diversity, and fairness. Furthermore, we analyze these impacts in a dynamic simulation environment that incorporates feedback loops, allowing us to observe how male and female users are differently affected over time. This comprehensive approach contributes to the development of more equitable and user-sensitive recommendation strategies.

3 Materials and methods

This section explains the feedback loop, experimental methodology, dataset, algorithms, and evaluation metrics under separate subheadings.

3.1 Feedback loop

This section presents the simulation developed to model feedback loops in RSs. The simulation is designed to analyze the dynamic interactions between recommendation algorithms and user profiles over multiple iterations, emphasizing the impact of popularity bias, fairness, and beyond-accuracy metrics on various user segments.

3.1.1 Initialize rating matrix

The process starts with the user-item rating matrix R, where each entry $r_{u,i}$ shows the interaction between user u and item i. This matrix serves as the basis for predicting user preferences and generating top-*N* recommendations.

3.1.2 Categorize items into popular and niche

Items are sorted by rating frequency (f_i) in descending order. Pareto principle, the cumulative top 20% of items that account for approximately 80% of all ratings are classified as Popular (Head) items, while the remaining items are considered Niche (Tail). This classification enables a more realistic and balanced analysis of the impact of recommendation algorithms on item exposure.

3.1.3 Group users by gender

Users are divided into two groups, Women and Men, using the gender information from the dataset. This segmentation supports a targeted analysis of recommendation dynamics for different demographic groups.

3.1.4 Feedback loop iterations

The framework operates through an iterative feedback loop, consisting of the following sequential steps:

- *Compute Predictions:* A CF algorithm is used to estimate the predicted rating $\hat{r}_{u,i}$ for each user-item pair (u, i) in the dataset.
- *Generate top-N Recommendations*: For each user, a personalized top-*N* recommendation list is created by ranking items in descending order based on their predicted ratings.
- *Evaluate Recommendations*: The quality of the generated recommendations is assessed using a combination of accuracy, fairness, and beyond-accuracy metrics (e.g., diversity, novelty, calibration). To analyze group-specific impacts, metrics are aggregated separately for each gender group by averaging individual user scores.
- *Update User Profiles*: To simulate user interactions within the feedback loop, synthetic ratings are added to user profiles in a controlled and consistent manner. First, each user's profile size is computed and normalized using *min-max* scaling to ensure equitable treatment across users with varying activity levels. The normalized profile size determines the number of synthetic interactions c_u to be added from the current top-*N* recommendation list. Next, each user's historical average rating μ_u is calculated and used as the synthetic rating value. This rating is then assigned to the top-c_u items in the recommendation list, simulating the user's engagement with the most highly ranked items. Note that this simulation assumes full engagement with the top-c_u items in each iteration. Future enhancements may include modeling stochastic behavior such as partial interaction or user dropout to better reflect real-world dynamics.
- *Matrix Update and Iteration Continuation:* After each iteration, the user-item matrix is updated to include these synthetic interactions, enabling the recommender system to model how user profiles evolve over multiple iterations. This procedure ensures realistic feedback loop dynamics by preserving consistency with users' prior rating behavior while reflecting their personalized recommendation experience

3.2 Experimental setup and procedure

Recommendation lists were generated using *leave-one-out cross-validation*, where one interaction from each user's profile is used as the test set, while the remaining interactions form the training set. The recommendation model is trained on this training set, and predictions are made for all items concerning the held-out interaction. This process is repeated for every user, ensuring a thorough evaluation of each one. The top-*N* items with the highest prediction scores are then selected for recommendation, with N =10 in this study.

Experiments were conducted using the proposed framework over 10 iterations, dynamically updating user profiles to simulate a feedback loop. Metrics were calculated for the Women and Men groups to analyze changes in recommendation quality, fairness, and beyond-accuracy dimensions over time. This iterative approach effectively captures the dynamic effects of popularity bias and evaluates algorithm performance in adapting to gender-based user preferences.

3.3 Dataset

This study utilizes the MovieLens-1M (ML) dataset, which comprises 1,000,209 ratings from 6,040 users across

3,900 movies [31]. It provides rich demographic information, including age, gender, and occupation, alongside movie metadata such as genres and release years [32]. This detailed information supports a comprehensive analysis of recommendation dynamics and popularity bias.

3.4 Used CF algorithms

This study utilizes three advanced CF algorithms: Hierarchical Poisson Factorization (HPF) [33], Maximum Margin Matrix Factorization (MMMF) [34], and Variational Autoencoder for Collaborative Filtering (VAECF) [35].

These algorithms utilize various modeling techniques to enhance both recommendation accuracy and diversity. HPF is a probabilistic model that represents user preferences using latent factors, modeling user-item interactions with a Poisson distribution. This approach effectively handles sparse data and cold-start users by leveraging hierarchical priors. MMMF, on the other hand, optimizes item rankings by maximizing the margin between relevant and irrelevant items. Unlike traditional matrix factorization models, MMMF focuses on preserving the relative order of items, making it particularly effective for ranking tasks. Meanwhile, VAECF employs deep generative models to learn complex interaction patterns by encoding user preferences into a probabilistic latent space and reconstructing them for rating prediction. This non-linear approach allows VAECF to capture intricate relationships between users and items, resulting in more accurate and diverse recommendations.

These algorithms were selected due to their diverse methodological foundations, including probabilistic, margin-based, and deep generative approaches, which provide a balanced comparative ground for analyzing feedback dynamics. Rather than aiming for state-of-the-art accuracy, this study prioritizes diversity in algorithmic perspectives to assess fairness and bias amplification over time.

Common algorithms, such as basic matrix factorization, BPR, or graph-based recommenders, were excluded to maintain focus and interpretability in long-term simulations. Including a broader range of models would increase computational complexity and reduce the clarity of longitudinal comparisons.

All algorithms were implemented using the Cornac framework [36] in Python, with hyperparameter settings aligned with the original publications to ensure reproducibility and consistency in experimental evaluations.

3.5 Evaluation metrics

This study evaluates the impact of popularity bias within the proposed feedback loop framework using a diverse set of metrics, including accuracy, popularity bias, calibration, and beyond-accuracy measures [21, 37, 38]. These metrics collectively assess the quality, fairness, and diversity of recommendations, providing a comprehensive evaluation of system performance from multiple perspectives.

Fairness metrics assess whether a recommendation system provides balanced outcomes for both users and items. From the user perspective, they measure alignment with individual preferences while mitigating biases that may disadvantage certain groups. From the item perspective, they ensure fair exposure across different popularity levels, preventing an overemphasis on popular content. This study employs multiple fairness metrics to evaluate both user- and item-based fairness dimensions.

3.5.1 Group Average Popularity (ΔGAP)

This metric assesses how recommendation algorithms impact item popularity across various user groups [11]. It compares the average popularity of items in users' historical profiles ($GAP_p(g)$) with the popularity of recommended items ($GAP_r(g)$). A smaller ΔGAP indicates better alignment between recommendations and users' past preferences, promoting fairness.

The metric is computed as:

$$\Delta GAP = \frac{GAPr(g) - GAPp(g)}{GAPp(g)} \tag{1}$$

A positive ΔGAP suggests that recommended items are more popular than users' historical preferences, while a negative ΔGAP indicates recommendations favor niche items. A value of zero reflects a fair balance.

3.5.2 Mean Rank Miscalibration (MRMC)

MRMC measures fairness from a user perspective, evaluating how well recommendations align with a user's historical preferences for popular (*Head*) and niche (*Tail*) items [39]. It calculates the divergence between the user's historical popularity distribution (p) and the cumulative popularity distribution of recommended items at each rank j.

For a given user u, Rank Miscalibration (RMC) is computed as:

$$RMC(u) = \frac{\sum_{j=1}^{N} \text{Divergence}\left(p, q\left(R_j^*\right)\right)}{N} \tag{2}$$

The overall MRMC score is obtained by averaging RMC across all users:

$$MRMC = \frac{\sum_{u \in U} RMC(u)}{|U|} \tag{3}$$

A lower MRMC value indicates better alignment between recommendations and user preferences, ensuring fairer and more personalized suggestions.

3.5.3 Average Popularity of the Recommended Items (APRI)

The APRI metric assesses the popularity bias in recommendation lists by measuring the average popularity of the recommended items. It helps evaluate whether the system disproportionately favors widely known items over less popular ones [21].

For a given top-N recommendation list $\{i_1, i_2, ..., i_n\}$ for user u, each item's popularity (P_i) is calculated as the

proportion of users who have rated that item in the dataset. The APRI score is then computed as:

$$APRI = \sum_{i \in N} \frac{P_i}{|N|} \tag{4}$$

where N is the set of recommended items, and P_i represents an item's popularity. A lower APRI score indicates a fairer recommendation system by ensuring a more balanced exposure of items with different popularity levels.

3.5.4 Ratio of Popular Items (RPI)

The RPI metric measures the proportion of popular items in the top-N recommendation list, differentiating between frequently rated (head) and less popular (tail) items [21]. Unlike APRI, which evaluates general popularity levels, RPI focuses on the categorical distribution of recommended items.

Popular items are identified using the Pareto principle [40], where "head" items account for 20% of total ratings. Given a top-N recommendation list $\{i_1, i_2, ..., i_n\}$, the RPI score is calculated as:

$$RPI = \frac{\sum_{i \in N} \mathbb{1}(i \in H)}{|N|} \tag{5}$$

where H represents the set of head items, and $\mathbb{1}(i \in H)$ is an indicator function returning 1 if item i belongs to H and 0 otherwise.

A lower RPI score indicates a more diverse recommendation list, reducing the dominance of popular items and promoting the inclusion of niche content.

3.5.5 Normalized Discounted Cumulative Gain (nDCG)

The nDCG metric assesses the effectiveness of a recommendation system in ranking items according to user preferences [41]. It accounts for both the position of recommended items and their actual relevance, ensuring that higher-rated items appear earlier in the list.

To achieve this, the metric first computes the Discounted Cumulative Gain (DCG) by summing the relevance scores of recommended items, where relevance is discounted logarithmically based on rank. Then, the Ideal DCG (IDCG) is calculated by ranking items in the best possible order according to their actual ratings. The final nDCG score is obtained by normalizing DCG with IDCG, ensuring that scores range between 0 and 1. The formula for nDCG is given in Equation (6):

$$nDCG_u^N = \frac{DCG_u^N}{IDCG_u^N} \tag{6}$$

where DCG measures the gain from ranked items, and IDCG represents the maximum possible gain if items were perfectly ordered. A higher nDCG score indicates that the system ranks relevant items more effectively, improving recommendation accuracy.

3.5.6 Precision, Recall, and F1-Score

To assess the accuracy of top-N recommendation lists, we use *Precision*, *Recall*, and *F1-score*, which collectively evaluate how well the system delivers relevant recommendations.

Precision ($P@N_u$) measures the proportion of recommended items that are actually relevant to the user, assessing recommendation accuracy. *Recall* ($R@N_u$) calculates the proportion of relevant items successfully retrieved out of all possible relevant items in the user's profile. To determine relevance, items rated *4* or *5* on a 5-star scale are considered suitable [42].

The *F1-score* ($F1@N_u$) balances precision and recall, providing a harmonic mean of both metrics, as defined in Equation (7):

$$F1@N_u = 2 \times \frac{P@N_u \times R@N_u}{P@N_u + R@N_u} \tag{7}$$

A higher *F1-score* indicates a well-balanced recommendation list that optimally captures both relevance and coverage.

3.5.7 Average Percentage of Long-tail Items (APLT)

The APLT metric measures the proportion of recommended items that belong to the long-tail portion of the catalog, promoting diversity and reducing popularity bias [24]. Based on the Pareto principle [40], items are categorized as "head" (top 20% of items receiving 80% of ratings) or "tail" (remaining items). The APLT score is calculated as shown in Equation (8):

$$APLT_u = \frac{|\{i \mid i \in (N_u \cap T)\}|}{|N_u|} \tag{8}$$

where N_u is the set of recommended items for user u, and T represents long-tail items. A higher APLT value indicates a greater emphasis on less popular items, enhancing recommendation diversity.

3.5.8 Novelty

The Novelty metric assesses a recommendation system's ability to introduce users to new items rather than repeatedly suggesting familiar ones [43]. It measures the proportion of recommended items that the user has not previously rated, encouraging content exploration. The Novelty score is computed as shown in Equation (9):

$$\text{Novelty}_u = \frac{|\{i \mid i \notin I_u\}|}{|N_u|} \tag{9}$$

where N_u is the set of recommended items, and I_u represents items the user has already rated. A higher Novelty value indicates that the system effectively diversifies recommendations by suggesting previously unseen content.

3.5.9 Entropy

The Entropy metric quantifies the diversity of recommended items by assessing how evenly they are

distributed across the catalog [44]. It evaluates whether the system over-recommends a small subset of items or provides a more balanced selection. Entropy is computed using Equation (10):

$$\text{Entropy} = -\sum_{i \in K} \Pr(i) \log_2 \Pr(i) \tag{10}$$

where $\Pr(i)$ represents the relative frequency of item i in the combined recommendation lists. A higher Entropy value indicates greater diversity, ensuring a broader range of items are recommended.

3.5.10 Long Tail Coverge (LTC)

The LTC metric measures how well a recommendation system incorporates long-tail items, promoting diversity and fairness [24]. It evaluates whether the system effectively recommends less popular items instead of focusing solely on the most popular ones. The LTC score is calculated as shown in Equation (11):

$$LTC = \frac{|I_{\mathbb{N} \cap T}|}{|T|} \tag{11}$$

where $|T|$ represents the total number of long-tail items, and $|I_{\mathbb{N} \cap T}|$ is the set of recommended long-tail items. A higher LTC score indicates improved long-tail representation, enhancing recommendation diversity.

4 Experimental results

This section provides a comprehensive evaluation of the proposed framework by analyzing the performance of the HPF, MMMF, and VAECF algorithms across the MLM and PER datasets over $t = 10$ iterations. The results are assessed using fairness, accuracy, and beyond-accuracy metrics. The initial iteration results are examined, and the proportional changes between the 1st and 10th iterations are analyzed. The primary objective of these analyses is to understand how the recommendation framework differently impacts male and female user groups. The focus is particularly on how the iterative feedback loop influences the metrics. This approach enables a deeper understanding of the dynamics introduced by the simulation and the varying effects on different user profiles. Additionally, paired *t*-tests were conducted to assess whether the observed differences between different gender groups were statistically significant.

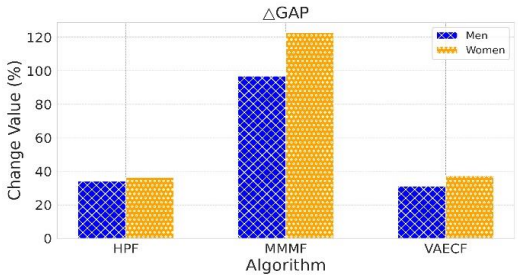

Figure 1. Proportional changes in ΔGAP from the 1st to the 10th iteration for different gender groups in the MLM dataset.

Figure 1 illustrates the proportional changes in the ΔGAP metric between the 1st and 10th iterations for different gender groups across the HPF, MMMF, and VAECF algorithms. Examining the initial iteration values, the HPF algorithm recorded 160.10 for men and 152.37 for women, while the VAECF algorithm showed 103.27 for men and 96.47 for women. The MMMF algorithm had the lowest initial ΔGAP values, with 61.54 for men and 49.78 for women.

When these initial values are considered alongside the proportional changes throughout the iterations, it is observed that the HPF algorithm, despite having the highest initial ΔGAP values, exhibited relatively lower proportional changes over the iterations. Similarly, the VAECF algorithm, which started with lower ΔGAP values compared to HPF, did not demonstrate significant changes throughout the iterations. On the other hand, the MMMF algorithm, which initially had the lowest ΔGAP values, showed the most substantial proportional change over the iterations. This suggests that the MMMF algorithm optimized the model more aggressively during the learning process and had a greater impact on gender-based differences as iterations progressed. Notably, the changes observed in the MMMF algorithm were more pronounced for women compared to men, indicating that the model may exhibit different learning dynamics based on gender. Furthermore, the differences observed for the MMMF and VAECF algorithms were found to be statistically significant at the 99% confidence level.

Figure 2 illustrates the proportional changes in the MRMC metric between the 1st and 10th iterations for different gender groups across the HPF, MMMF, and VAECF algorithms. Examining the initial iteration values, the HPF algorithm recorded an initial MRMC value of 0.3149 for men and 0.3618 for women. For the VAECF algorithm, these values were 0.2310 for men and 0.2437 for women, while the MMMF algorithm had the lowest initial MRMC values, with 0.1381 for men and 0.1314 for women.

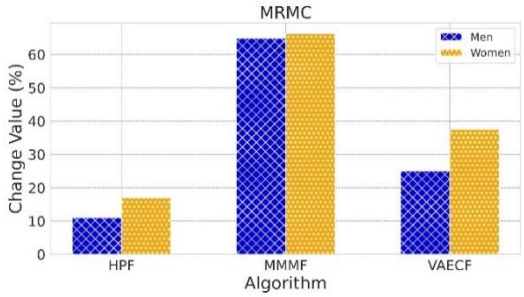

Figure 2. Proportional changes in MRMC from the 1st to the 10th iteration for different gender groups in the MLM dataset.

When these initial values are considered alongside the proportional changes throughout the iterations, it is observed that the HPF algorithm, despite having the highest initial MRMC values, exhibited the lowest proportional change over the iterations. Similarly, the VAECF algorithm, which started with a lower MRMC value compared to HPF,

demonstrated a more noticeable change throughout the iterations. On the other hand, the MMMF algorithm, which initially had the lowest MRMC values, showed the most substantial proportional change, indicating a greater impact on model optimization. Additionally, the MMMF algorithm exhibited a highly similar change for both men and women, suggesting a more balanced learning process between gender groups compared to the other algorithms. In contrast, the VAECF algorithm showed a higher proportional change for women than for men, indicating that this algorithm may exhibit different learning dynamics based on gender. These findings suggest that the initial MRMC levels are directly related to the rate of change in optimization during the iteration process and that the model's learning dynamics may vary across gender groups as iterations progress. Moreover, the differences observed for the HPF and VAECF algorithms were found to be statistically significant at the 99% confidence level.

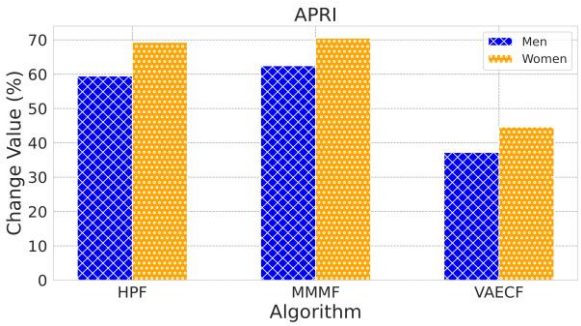

Figure 3. Proportional changes in APRI from the 1st to the 10th iteration for different gender groups in the MLM dataset.

Figure 3 illustrates the proportional changes in the APRI metric between the 1st and 10th iterations for different gender groups across the HPF, MMMF, and VAECF algorithms. Examining the initial iteration values, the HPF algorithm had the highest starting APRI values (men: 0.3912, women: 0.3638), the VAECF algorithm showed moderate values (men: 0.3229, women: 0.2985), and the MMMF algorithm had the lowest initial APRI levels (men: 0.2507, women: 0.2219).

When these initial values are considered alongside the proportional changes throughout the iterations, it is observed that the MMMF algorithm, despite having the lowest initial APRI values, exhibited the highest proportional change. In contrast, the HPF and VAECF algorithms showed more limited variations, with MMMF demonstrating the most significant change for both men and women. This suggests that the MMMF algorithm had a greater impact on the model's optimization process and that there are notable differences in learning dynamics among the algorithms.

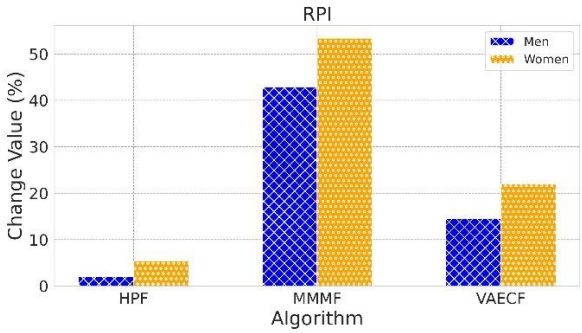

Figure 4. Proportional changes in RPI from the 1st to the 10th iteration for different gender groups in the MLM dataset.

Figure 4 illustrates the proportional changes in the RPI metric between the 1st and 10th iterations for different gender groups across the HPF, MMMF, and VAECF algorithms. Examining the initial iteration values, the HPF algorithm had the highest starting RPI values (men: 0.9606, women: 0.9145), the VAECF algorithm showed moderate values (men: 0.7722, women: 0.6985), and the MMMF algorithm had the lowest initial RPI levels (men: 0.5214, women: 0.4275).

When these initial values are considered alongside the proportional changes throughout the iterations, it is observed that the MMMF algorithm, despite having the lowest initial RPI values, exhibited the highest proportional change. In contrast, the HPF and VAECF algorithms showed more limited variations, with MMMF demonstrating a more significant change for women. This suggests that the MMMF algorithm had a greater impact on the model's optimization process and that gender-based differences in learning dynamics may exist.

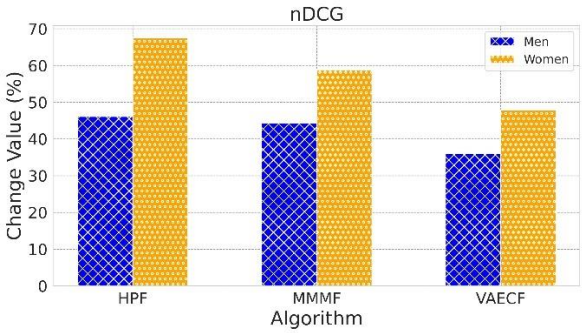

Figure 5. Proportional changes in nDCG from the 1st to the 10th iteration for different gender groups in the MLM dataset.

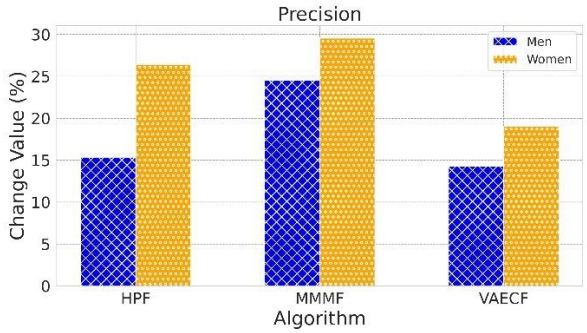

Figure 6. Proportional changes in Precision from the 1st to the 10th iteration for different gender groups in the MLM dataset.

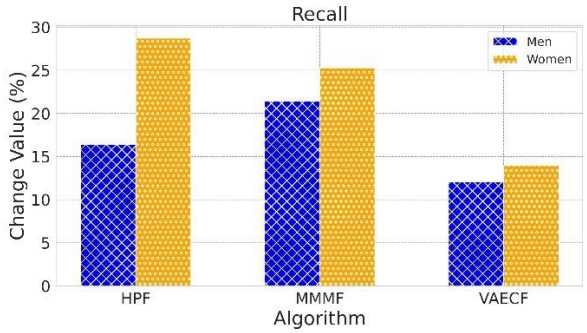

Figure 7. Proportional changes in Recall from the 1st to the 10th iteration for different gender groups in the MLM dataset.

The evaluation of the HPF, MMMF, and VAECF algorithms across various accuracy metrics, including nDCG, Precision, Recall, and F1-Score, reveals consistent patterns in performance and improvement trends. Analyzing the initial iteration values, it is observed that the VAECF algorithm has the highest starting values across all metrics for both male and female users. The HPF algorithm ranks second in initial performance, while the MMMF algorithm exhibits the lowest baseline levels. These findings indicate that VAECF is initially more effective; however, differences in improvement trends throughout the iteration process become a key factor in distinguishing the algorithms.

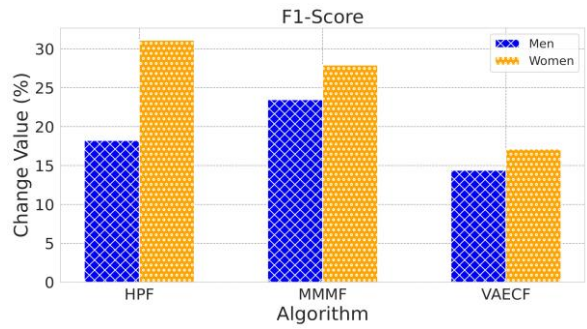

Figure 8. Proportional changes in F1-Score from the 1st to the 10th iteration for different gender groups in the MLM dataset.

When evaluating the proportional changes between the 1st and 10th iterations, significant improvements in all accuracy metrics are observed for the MMMF and HPF algorithms, and these improvements are reflected in Figures 5, 6, 7, and 8. The HPF algorithm demonstrates a more pronounced performance increase, particularly for female users, while the MMMF algorithm shows notable improvement across accuracy metrics, maintaining a more balanced performance enhancement across genders. In contrast, despite having the highest initial values, the VAECF algorithm exhibits a more limited change throughout the iteration process compared to the other algorithms. This suggests that HPF and MMMF have greater adaptability and optimization potential, resulting in significant performance improvements over iterations. Furthermore, differences in learning dynamics between genders are evident, as HPF and MMMF offer higher proportional improvements for female users, while VAECF maintains a more stable progression. These findings indicate that although VAECF provides a strong initial predictive capability, HPF and MMMF offer greater optimization advantages throughout the iterative learning process. Additionally, the differences observed in the APRI, RPI, nDCG, Precision, Recall, and F1-score metrics between male and female user groups were found to be statistically significant for all algorithms at the 99% confidence level.

The evaluation of the HPF, MMMF, and VAECF algorithms on the APLT metric highlights notable differences in computational efficiency. At the first iteration, MMMF has the highest APLT values, with 0.4786 for males and 0.5725 for females, indicating that it requires the longest prediction time. VAECF follows with 0.2278 for males and 0.3015 for females, while HPF demonstrates the lowest initial latency, with 0.0394 for males and 0.0855 for females, suggesting a more efficient computational process.

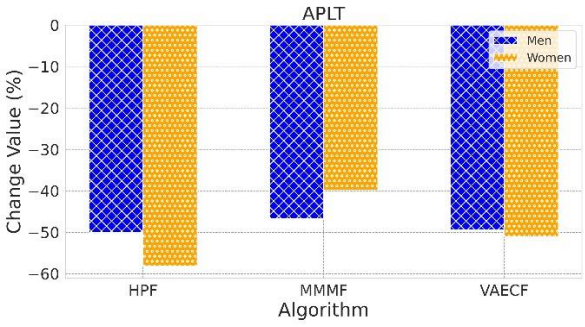

Figure 9. Proportional changes in APLT from the 1st to the 10th iteration for different gender groups in the MLM dataset.

As shown in Figure 9, the proportional changes between the 1st and 10th iterations reveal a substantial reduction in APLT for all algorithms. HPF and VAECF experience the most significant decrease, particularly for female users, where VAECF shows a steeper decline. The MMMF algorithm, despite having the highest initial APLT values, also undergoes a considerable reduction. However, it still maintains relatively higher latency compared to HPF and

VAECF. Also, the differences observed in the APLT metric between male and female user groups were found to be statistically significant when using the HPF and MMMF algorithms.

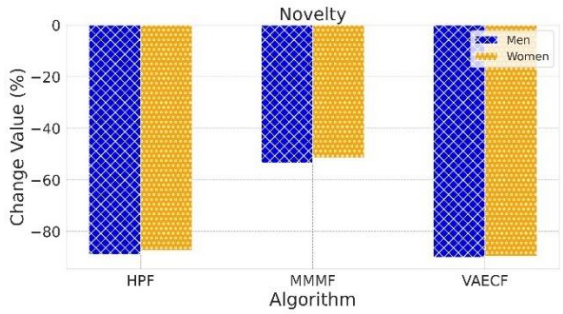

Figure 10. Proportional changes in Novelty from the 1st to the 10th iteration for different gender groups in the MLM dataset.

The evaluation of the HPF, MMMF, and VAECF algorithms on the novelty metric highlights significant differences in recommendation diversity. Analyzing the initial iteration values, the MMMF algorithm has the highest starting *novelty* levels, with 0.4387 for males and 0.5099 for females. The HPF algorithm yields values of 0.3998 for males and 0.4891 for females, while the VAECF algorithm has the lowest initial values, at 0.3198 for males and 0.3863 for females. This suggests that the MMMF and HPF algorithms initially provide a broader recommendation set, whereas VAECF generates more restricted suggestions.

As shown in Figure 10, the proportional changes between the 1st and 10th iterations indicate a substantial decline in *novelty* values for all algorithms. HPF and VAECF, in particular, show a significantly larger decrease for female users. Although MMMF starts with the highest *novelty* levels, it experiences a more gradual decline compared to the other algorithms. This indicates that HPF and VAECF refine their recommendations more aggressively over iterations, leading to a more specialized set of suggested items. In contrast, MMMF maintains a relatively broader recommendation space with a more balanced reduction in novelty. Also, the differences observed in the novelty metric between male and female user groups were not found to be statistically significant for any of the algorithms.

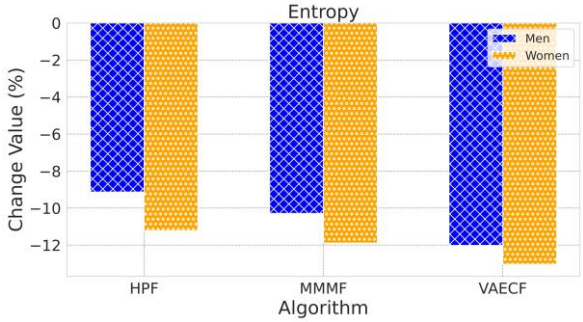

Figure 11. Proportional changes in Entropy from the 1st to the 10th iteration for different gender groups in the MLM dataset.

The evaluation of the HPF, MMMF, and VAECF algorithms on the entropy metric highlights differences in the uncertainty and diversity of recommendation systems. Analyzing the initial iteration values, the VAECF algorithm has the highest starting entropy levels, with 0.0001391 for males and 0.0003552 for females. The MMMF algorithm follows, with values of 0.0001290 for males and 0.0003317 for females. The HPF algorithm has the lowest initial entropy values, at 0.0001036 for males and 0.0002704 for females, indicating that its recommendation system is the least uncertain.

As shown in Figure 11, the proportional changes between the 1st and 10th iterations indicate a decline in entropy values for all algorithms. This reduction is more pronounced for female users. Despite starting with the highest entropy values, the VAECF algorithm experiences the sharpest decline throughout the iterations. The MMMF algorithm also shows a noticeable reduction, though less drastic than VAECF. The HPF algorithm, while initially having the lowest entropy levels, maintains a more balanced decrease over the iterations. This suggests that HPF gradually reduces recommendation diversity in a controlled manner, whereas VAECF undergoes a more drastic shift toward a narrower range of recommendations. Additionally, the differences observed in entropy between male and female user groups were found to be statistically significant when using the HPF and MMMF algorithms.

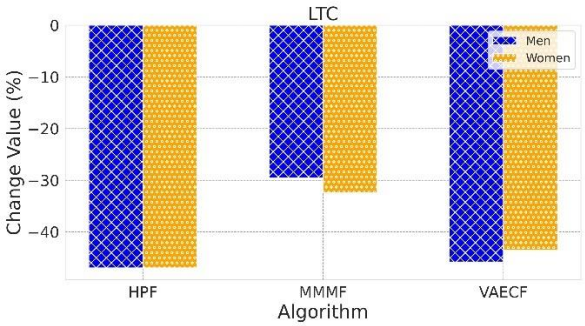

Figure 12. Proportional changes in LTC from the 1st to the 10th iteration for different gender groups in the MLM dataset.

The evaluation of the HPF, MMMF, and VAECF algorithms on the LTC metric highlights differences in how recommendation systems adapt to user preferences. Analyzing the initial iteration values, the VAECF algorithm has the highest starting LTC levels, with 3.36206E-05 for males and 6.96556E-05 for females. The MMMF algorithm yields results of 2.4599E-05 for males and 5.30419E-05 for females. The HPF algorithm has the lowest initial LTC values, at 9.02163E-07 for males and 2.28629E-06 for females, indicating that it has the slowest learning rate in terms of recommendation adjustments.

As shown in Figure 12, the proportional changes between the 1st and 10th iterations indicate a significant decline in LTC values for all algorithms. HPF and VAECF, in particular, show a more substantial reduction for female

users. Although VAECF starts with the highest LTC values, it experiences the sharpest decline throughout the iterations. The MMMF algorithm also shows a noticeable reduction, though less drastic than VAECF. The HPF algorithm, while having the lowest initial levels, maintains the most balanced decrease over iterations. These findings suggest that HPF optimizes its recommendation system more steadily, while VAECF, despite its initially rapid learning, undergoes the most significant shift in later iterations. The differences observed in the LTC metric between male and female user groups were found to be statistically significant at the 99% confidence level only when using the MMMF algorithm.

5 Limitations

This study assumes that users consume all items recommended in the top-*N* lists at each iteration, which simplifies the simulation of the feedback loop. However, in real-world scenarios, users rarely consume the entire recommendation list, and the proportion of consumed items can vary widely across different domains and user groups. Importantly, consumption patterns may exhibit gender-related differences, with male and female users potentially engaging with recommended content in different ways.

This limitation could impact the observed dynamics of fairness, popularity bias, and diversity in the system. For instance, if one gender tends to consume fewer recommendations, the feedback loops and resulting biases might evolve differently than modeled here. Furthermore, because the simulation generates fully synthetic user interactions, the findings may not fully capture complex real-world behaviors such as partial engagement, user fatigue, or changing preferences over time. Therefore, caution should be exercised when generalizing the results, and future work should validate the framework using real user data or more sophisticated behavioral models.

Therefore, incorporating variable consumption rates, potentially conditioned on demographic factors such as gender, would provide a more realistic simulation framework and deeper insights into the dynamics of fairness.

6 Conclusion and future work

This study proposes a dynamic feedback loop framework to analyze the long-term impacts of gender-based fairness and popularity bias in recommender systems. The developed simulation infrastructure models the system's evolution over time by updating user profiles at each iteration based on the recommended items. This framework enables the tracking of how fairness metrics change dynamically as synthetic feedback accumulates over time, thereby filling a gap in fairness research under feedback loop conditions.

Three collaborative filtering algorithms (HPF, MMMF, and VAECF) were evaluated in terms of both their initial performance and their progression over multiple iterations of the feedback loop. The results indicate that the MMMF algorithm achieved higher values for calibration and diversity metrics among female users. The HPF algorithm demonstrated more balanced improvement in overall accuracy metrics, whereas the VAECF algorithm initially showed strong performance but limited gains as the feedback loop progressed.

In terms of fairness metrics, popularity-driven disparities—measured using indicators such as GAP, MRMC, and RPI—increased over time. Higher values in these metrics were observed for the female user group, highlighting the growing differentiation across user groups as iterations progressed. Additionally, the system faced increasing difficulty in maintaining access to long-tail content and preserving diversity over time. Our proportional change analysis across iterations uncovered gender-specific learning behaviors, offering actionable insights for fairness-aware algorithm design.

These findings suggest that recommender systems should not only focus on initial accuracy performance but also monitor how fairness, diversity, and calibration evolve dynamically throughout user interaction. The differing impacts on user groups, specifically those based on gender in this study, underscore the necessity of designing more adaptive, user-aware algorithms. This study provides the first longitudinal, gender-specific comparison of HPF, MMMF, and VAECF in this context, revealing how feedback loops can intensify disparities in both fairness and performance metrics.

For future research, incorporating stochastic models of user interaction, such as varying engagement levels, feedback noise, and dropout behaviors, could enhance the realism of simulations. Additionally, extending the demographic analysis beyond gender (e.g., age, occupation) and validating the framework on larger, real-world datasets will contribute to more generalizable insights. Developing personalized and group-sensitive recommendation strategies will be essential for mitigating fairness concerns that arise in feedback-driven systems.

Acknowledgements

This work is based on the Master's thesis of Yildiz Zoralioglu and builds upon the research conducted as part of her graduate studies.

Declaration of AI Statement

During the preparation of this manuscript, the authors used ChatGPT to improve the grammar, readability, and fluency of the text. All content was subsequently reviewed by the authors, who take full responsibility for the final version of the work.

Conflict of Interest

The authors declare that they have no conflict of interest.

Similarity rate (iThenticate): %14

References

[1] G. Adomavicius and A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17 (6), 734–749, 2005. https://doi.org/10.1109/TKDE.2005.99.

[2] F. Ricci, L. Rokach and B. Shapira, Recommender Systems Handbook. Springer, 2015.

[3] X. Su and T. M. Khoshgoftaar, A survey of collaborative filtering techniques. Advances in Artificial Intelligence, 2009, Article ID 421425, 1–19, 2009. https://doi.org/10.1155/2009/421425.

[4] Y. Koren, R. Bell and C. Volinsky, Matrix factorization techniques for recommender systems. Computer, 42 (8), 30–37, 2009. https://doi.org/10.1109/MC.2009.263.

[5] G. Linden, B. Smith and J. York, Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing, 7 (1), 76–80, 2003. https://doi.org/10.1109/MIC.2003.1167344.

[6] G. Adomavicius and Y. Kwon, Improving aggregate recommendation diversity using ranking-based techniques. IEEE Transactions on Knowledge and Data Engineering, 24 (5), 896–911, 2011. https://doi.org/10.1109/TKDE.2011.15.

[7] H. Abdollahpouri, M. Mansoury, R. Burke and B. Mobasher, The unfairness of popularity bias in recommendation. Proceedings of the Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML), 1–5, Long Beach, CA, USA, 2019.

[8] L. Boratto, G. Fenu and M. Marras, Connecting user and item perspectives in popularity debiasing for collaborative recommendation. Information Processing & Management, 58 (1), 102387, 2021. https://doi.org/10.1016/j.ipm.2020.102387.

[9] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher and E. Malthouse, User-centered evaluation of popularity bias in recommender systems. Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21), 119–128, Utrecht, Netherlands, 2021. https://doi.org/10.1145/3450613.3456821.

[10] R. Sinha and K. Swearingen, Comparing recommendations made by online systems and by friends. DELOS-NSF Workshop on Personalization and Recommender Systems in Digital Libraries, 1–10, Dublin, Ireland, 2001.

[11] H. Abdollahpouri, R. Burke and B. Mobasher, Managing popularity bias in recommender systems with personalized re-ranking. Proceedings of the 32nd International Florida Artificial Intelligence Research Society Conference (FLAIRS), 413–418, Sarasota, FL, USA, 2019.

[12] D. Turnbull, S. McQuillan, S. Zhang and D. Morrison, Exploring popularity bias in music recommendation models and commercial systems. arXiv preprint, arXiv:2208.09517, 2022. https://doi.org/10.48550/arXiv.2208.09517.

[13] H. Abdollahpouri, M. Mansoury, R. Burke and B. Mobasher, The connection between popularity bias, calibration, and fairness in recommendation. Proceedings of the 14th ACM Conference on Recommender Systems (RecSys '20), 726–731, Virtual Event, Brazil, 2020. https://doi.org/10.1145/3383313.3418487.

[14] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas and F. Diaz, Towards a fair marketplace: Counterfactual evaluation of the fairness of exposure in recommender systems. Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18), 2243–2251, 2018. https://doi.org/10.1145/3269206.3272027.

[15] D. Kowald, M. Schedl and E. Lex, The unfairness of popularity bias in music recommendation: A reproducibility study. Advances in Information Retrieval. ECIR 2020. Lecture Notes in Computer Science, 12036, 35–42, 2020. https://doi.org/10.1007/978-3-030-45442-5_5.

[16] M. D. Ekstrand, M. Tian, I. M. Azpiazu, J. D. Ekstrand, O. Anuyah, D. McNeill and M. S. Pera, All the cool kids, how do they fit in? Popularity and demographic biases in recommender evaluation and effectiveness. Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT '18), 172–186, New York, NY, USA, 2018.

[17] M. Gulsoy, E. Yalcin and A. Bilge, Robustness of privacy-preserving collaborative recommenders against popularity bias problem. PeerJ Computer Science, 9, e1438, 2023. https://doi.org/10.7717/peerj-cs.1438.

[18] B. Ferwerda, M. Schedl and M. Tkalčič, Personality and taxonomy preferences, gender, and age in music recommender systems. Personal and Ubiquitous Computing, 23, 801–813, 2019. https://doi.org/10.1007/s11042-019-7336-7.

[19] Q. Zhu, J. Wang and J. Caverlee, Fairness-aware personalized ranking recommendation via adversarial learning. arXiv preprint, arXiv:2103.07849, 2021. https://doi.org/10.48550/arXiv.2103.07849.

[20] E. Yalcin and A. Bilge, Investigating and counteracting popularity bias in group recommendations. Information Processing & Management, 58 (5), 102608, 2021. https://doi.org/10.1016/j.ipm.2021.102608.

[21] E. Yalcin and A. Bilge, Treating adverse effects of blockbuster bias on beyond-accuracy quality of personalized recommendations. Engineering Science and Technology, an International Journal, 33, 101083, 2022. https://doi.org/10.1016/j.jestch.2021.101083.

[22] M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti and E. H. Chi, Top-k off-policy correction for a REINFORCE recommender system. Proceedings of the 13th ACM International Conference on Web Search and Data Mining (WSDM '20), 456–464, Houston, TX, USA, 2020.

[23] S. Yao and B. Huang, Beyond parity: Fairness objectives for collaborative filtering. Advances in Neural Information Processing Systems, 2017.

[24] H. Abdollahpouri, Popularity bias in recommendation: A multi-stakeholder perspective. arXiv preprint, arXiv:2008.08551, 2020. https://doi.org/10.48550/arXiv.2008.08551.

[25] Y. Wang, W. Ma, M. Zhang, Y. Liu and S. Ma, A survey on the fairness of recommender systems. ACM

Transactions on Information Systems, 40 (3), 1–44, 2022. https://doi.org/10.1145/3547333.

[26] Y. Deldjoo, D. Jannach, A. Bellogin, A. Difonzo and D. Zanzonelli, Fairness in recommender systems: Research landscape and future directions. User Modeling and User-Adapted Interaction, 34, 59–108, 2023. https://doi.org/10.1007/s11257-023-09364-z.

[27] G. Alves, D. Jannach, R. F. de Souza and M. G. Manzato, User perception of fairness-calibrated recommendations. Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP '24), 113–122, Limassol, Cyprus, 2024. https://doi.org/10.1145/3627043.3659558.

[28] A. J. B. Chaney, B. M. Stewart and B. E. Engelhardt, How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18), 224–232, 2018. https://doi.org/10.1145/3240323.3240370.

[29] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher and R. Burke, Feedback loop and bias amplification in recommender systems. Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20), 2145–2148, Virtual Event, Ireland, 2020. https://doi.org/10.1145/3340531.3412152.

[30] K. Krauth, Y. Wang and M. I. Jordan, Breaking feedback loops in recommender systems with causal inference. Proceedings of the 39th International Conference on Machine Learning (ICML '22), Baltimore, MD, USA, 2022.

[31] S. K. Lam, J. L. Herlocker, J. A. Konstan and J. T. Riedl, Movielens 1M: Collaborative filtering dataset for personalized recommendations. Proceedings of SIGIR, 35–42, 2008.

[32] F. M. Harper and J. A. Konstan, The MovieLens datasets: History and context. ACM Transactions on Interactive Intelligent Systems (TiiS), 5 (4), 1–19, 2015. https://doi.org/10.1145/2827872.

[33] P. Gopalan, J. M. Hofman and D. M. Blei, Scalable recommendation with hierarchical Poisson factorization. Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI '15), 326–335, Amsterdam, Netherlands, 2015.

[34] N. Srebro and T. Jaakkola, Weighted low-rank approximations. Proceedings of the 20th International Conference on Machine Learning (ICML '03), 720–727, Washington, DC, USA, 2003.

[35] D. Liang, R. G. Krishnan, M. D. Hoffman and T. Jebara, Variational autoencoders for collaborative filtering. Proceedings of the 2018 World Wide Web Conference (WWW '18), 689–698, Lyon, France, 2018. https://doi.org/10.1145/3178876.3186150.

[36] A. Salah, Q.-T. Truong and H. W. Lauw, Cornac: A comparative framework for multimodal recommender systems. Journal of Machine Learning Research, 21 (95), 1–5, 2020.

[37] L. Boratto, G. Fenu, M. Marras and G. Medda, Consumer fairness in recommender systems: Contextualizing definitions and mitigations. Proceedings of the 44th European Conference on Information Retrieval (ECIR '22), 552–566, Stavanger, Norway, 2022.

[38] S. M. McNee, J. Riedl and J. A. Konstan, Being accurate is not enough: How accuracy metrics have hurt recommender systems. CHI '06 Extended Abstracts on Human Factors in Computing Systems, 1097–1101, Montréal, Canada, 2006. https://doi.org/10.1145/1125451.1125659.

[39] D. C. Silva, M. G. Manzato and F. A. Durão, Exploiting personalized calibration and metrics for fairness recommendation. Expert Systems with Applications, 181, 115112, 2021. https://doi.org/10.1016/j.eswa.2021.115112.

[40] R. Sanders, The Pareto principle: Its use and abuse. Journal of Services Marketing, 1 (2), 37–40, 1987.

[41] P. J. Chia, J. Tagliabue, F. Bianchi, C. He and B. Ko, Beyond NDCG: Behavioral testing of recommender systems with RecList. Companion Proceedings of the Web Conference 2022 (WWW '22), 99–104, Virtual Event, 2022.

[42] J. Bobadilla, F. Ortega, A. Hernando and J. Bernal, A collaborative filtering approach to mitigate the new user cold start problem. Knowledge-Based Systems, 26, 225–238, 2012. https://doi.org/10.1016/j.knosys.2011.07.021.

[43] S. Wang, M. Gong, H. Li and J. Yang, Multi-objective optimization for long tail recommendation. Knowledge-Based Systems, 104, 145–155, 2016. https://doi.org/10.1016/j.knosys.2016.04.018.

[44] M. Elahi, B. Ferwerda, M. Tkalčič, M. Schedl and F. Ricci, Investigating the impact of recommender systems on user-based and item-based popularity bias. Information Processing & Management, 58 (5), 102655, 2021. https://doi.org/10.1016/j.ipm.2021.102655.