Explaining Recommendation Fairness from a User/Item Perspective

JIE LI, RMIT University, Australia
YONGLI REN, RMIT University, Australia
MARK SANDERSON, RMIT University, Australia
KE DENG, RMIT University, Australia

Recommender systems play a crucial role in personalizing user experiences, yet ensuring fairness in their outcomes remains an elusive challenge. This work explores the impact of individual users or items on the fairness of recommender systems, thus addressing a significant knowledge gap in the field. We introduce an innovative approach called Adding-based Counterfactual Fairness Reasoning (*ACFR*), designed to elucidate recommendation fairness from the unique perspectives of users and items. Conventional methodologies, like erasing-based counterfactual analysis, pose limitations, particularly in modern recommender systems dealing with a large number of users and items. These traditional methods, by excluding specific users or items, risk disrupting the crucial relational structure central to collaborative filtering recommendations. In contrast, *ACFR* employs an adding-based counterfactual analysis, a unique strategy allowing us to consider potential, yet-to-happen user-item interactions. This strategy preserves the core user-item relational structure, while predicting future behaviors of users or items. The commonly-used feature-based counterfactual analysis, relying on gradient-based optimization to identify interference on each feature, is not directly applicable in our case. In the recommendation scenario we consider, only interactions between users and items are present during model training—no distinct features are involved. Consequently, the traditional mechanism proves impractical for identifying interference on these existing interactions. Our extensive experiments validate the superiority of *ACFR* over traditional baseline methods, demonstrating significant improvements in recommendation fairness on benchmark datasets. This work, therefore, provides a fresh perspective and a promising methodology for enhancing fairness in recommender systems.

CCS Concepts: • **Information systems** → **Recommender systems**.

Additional Key Words and Phrases: Explainable Fairness, Counterfactual Reasoning, Recommender Systems

1 INTRODUCTION

Recommender systems have become indispensable in navigating the vast digital landscape, tailoring content, products, and services to meet individual preferences [18, 22, 23, 37, 72]. While these systems are central to enhancing user experience, the challenge of ensuring fairness within such algorithms is increasingly recognized [3, 12, 14, 19, 34, 40, 47, 56, 57, 68, 74]. The concept of fairness in recommender systems extends beyond mere algorithmic accuracy; it involves a nuanced understanding of how biases can inadvertently be perpetuated through user and item interactions. Traditional fairness explanation methods often overlook the granularity of these interactions [17, 18, 62, 66, 67], thereby missing a comprehensive assessment of fairness within these complex systems.

Authors' addresses: Jie Li, hey.jieli@gmail.com, RMIT University, Melbourne, Victoria, Australia; Yongli Ren, yongli.ren@rmit.edu.au, RMIT University, Melbourne, Victoria, Australia; Mark Sanderson, mark.sanderson@rmit.edu.au, RMIT University, Melbourne, Victoria, Australia; Ke Deng, ke.deng@rmit.edu.au, RMIT University, Melbourne, Victoria, Australia.

ACM 1558-2868/2024/10-ART
https://doi.org/10.1145/3698877

This work introduces **A**dding-based **C**ounterfactual **F**airness **R**easoning (*ACFR*), a novel methodology designed to bridge this gap by analyzing fairness from the individual perspectives of users and items. Through counterfactual reasoning [18, 50] and the generation of explanation scores, *ACFR* seeks to illuminate the intricate dynamics of how specific users and items contribute to or detract from fairness in recommendations. This approach offers a significant enhancement over traditional methods, providing a detailed exploration of the relationships between fairness and user-item interactions.

The evolution of recommender systems has undeniably enhanced user engagement by tailoring experiences to individual preferences. Yet, this personalization comes with the challenge of maintaining fairness, a multifaceted concept involving the equitable distribution of recommendation benefits among users and visibility among items. The interplay between fairness and accuracy further complicates this challenge, as efforts to optimize one can inadvertently compromise the other [12, 14, 34, 47, 68]. Our investigation seeks not to solve this dilemma outright but to explore the nuanced relationship between fairness and the myriad interactions that animate the recommender systems landscape. Against this backdrop, our research is propelled by a set of inquiries aimed at unraveling the complexities of fairness in recommender systems:

(1) **Research Question 1:** How do individual user-item interactions shape and potentially distort the fairness of recommender systems?
(2) **Research Question 2:** How can we develop a methodological approach that emphasizes individual user-item interactions to better understand the sources of unfairness?
(3) **Research Question 3:** How does the Adding-based Counterfactual Fairness Reasoning (*ACFR*) approach address the identified shortcomings of traditional fairness explanation methodologies?
(4) **Research Question 4:** What empirical evidence supports the efficacy of the *ACFR* approach in explaining recommendation fairness compared to existing methods?

Following the outlined research questions, our study delves into the intricacies of fairness in recommender systems, proposing a novel approach to uncover and elucidate the factors influencing fairness through individual interactions. This effort represents a pivotal move towards a more refined understanding of fairness, offering insights crucial for the development of transparent and equitable recommendation algorithms. To illustrate the depth of the issues at hand:

(1) **Exploring User-Specific Influences**: Imagine a scenario where certain users consistently receive recommendations that misalign with their interests. An in-depth analysis of these users' interactions may reveal that their unique or niche preferences are overlooked by current recommendation algorithms, leading to recommendations that starkly diverge from their actual interests. Understanding these user-item interactions is essential for refining algorithms to better accommodate diverse user interests, ultimately contributing to a fairer system.
(2) **Investigating Item-Specific Influences**: Conversely, consider the case where a narrow selection of items dominates recommendations across a broad user spectrum, despite the existence of a varied and extensive item catalog. A detailed examination of these particular items and their interactions with users might uncover that the algorithm disproportionately favors them due to their broad appeal or high ratings. Identifying such patterns is crucial for adjusting the system to encourage a wider distribution of items, enhancing fairness in the process.

The primary takeaway here is that achieving fairness in recommender systems goes beyond simply analysing broad groups of users and items - it necessitates an understanding of individual user-item interactions. A detailed examination of these interactions fosters a profound understanding of the factors causing unfairness, ultimately enabling a more tailored approach to rectifying these issues. With this work, we aim to bridge the existing knowledge gap by scrutinizing whether specific users or items exert a positive or negative impact on the fairness of recommender systems. To facilitate this, we put forth a novel approach, termed as Adding-based Counterfactual

Fairness Reasoning (*ACFR*), that probes into the concept of fairness from the unique lens of individual users and items.

Traditional methodologies, like erasing-based counterfactual analysis [18], often fall short in fully addressing this concern, particularly in the context of contemporary recommender systems handling an expansive network of users and items. By excluding specific users or items, these erasing-based methods disrupt the fundamental relational structure, a critical component in collaborative filtering recommendations. In contrast, *ACFR* uses an adding-based counterfactual analysis. This method speculates on future user-item behaviors by imputing potential interactions that have not yet occurred, thereby preserving the integral structure of user-item relations.

Feature-based counterfactual analysis [18] is another prevalent methodology. However, its direct application may not be feasible in our case. These methods typically apply gradient-based optimization to identify the impact of each feature, thereby determining which features substantially influence recommendation fairness. The procedure introduces random interferences for each feature and treats these interfered features as inputs to the trained recommendation model. Following this, the method maintains the features constant and strives to retain recommendation accuracy by minimizing the discrepancy between predictions and ground truths (i.e., predictions and interactions), while iteratively updating the interference to optimize a fairness objective.

However, in the context of recommendations we consider, features may not exist during the recommendation model training process - only interactions between users and items do. Hence, we only investigate the influence of interactions of specific users or items in this work. Contrary to features, interactions act as the ground truth rather than model inputs, rendering the updating of interference on interactions via gradient-based methods impractical when striving to minimize the accuracy-based loss between predictions and ground truths while keeping the ground truths constant. Indeed, any alteration in interference on interactions will inevitably modify the ground truths, making it unfeasible to optimize an objective when the ground truths are continuously changing. In order to address this issue, our approach employs a unique gradient-based method through our *ACFR* system. This method identifies the interferences on the imputed interactions between users and items to shed light on fairness. In essence, the *ACFR* framework enhances fairness in recommender systems by imputing hypothetical user-item interactions, simulating their potential impacts on fairness. It assesses each user and item's effect on fairness, identifying both positive and negative influences through a unique gradient-based optimization for imputation. This process provides actionable insights into fairness improvements.

At the heart of *ACFR* lies the intuitive principle of counterfactual reasoning: by imagining alternative scenarios 'what-ifs' where users interact with items differently, we can anticipate the potential impacts on fairness. This core intuition enables us to dissect the complex dynamics of recommender systems, identifying specific interactions that contribute to or detract from fairness. By reimagining the interaction matrix through the lens of these hypothetical scenarios, *ACFR* aims to not only illuminate the paths through which biases propagate but also guide the development of more equitable recommendation algorithms.

We will detail this methodology in section 4, highlighting its proven superiority over baseline methods through extensive experimentation. Through rigorous experimentation, we have demonstrated the superiority of our proposed *ACFR* solution over baseline methods, significantly enhancing recommendation fairness on benchmark datasets.

The main contributions of this work are as follows:

- This work, to the best of our knowledge, is the first effort on explainable recommendation fairness from the perspective of individual users and items based on their current user-item interactions.
- This work proposed a novel adding-based counterfactual analysis that is tailored to suit the uniqueness of our problem to overcome the shortcoming of the existing counterfactual-analysis-based method.

- This work verified the superiority of the proposed solution compared to baseline methods on the benchmark datasets by comprehensive experiments in various aspects. The source code of this work is publicly available via the link[1].

2 RELATED WORK

2.1 Fairness in Ranking and Recommendations

Current recommender systems serve as intermediaries on multi-sided platforms by providing recommendations that connect users to items. Based on this context, fairness concerns in recommender systems have drawn more attention. Fairness in recommendation can be divided into user fairness [17, 33, 35, 47] and item fairness [34, 40, 43, 55, 71]. For example, [35] considers the user perspective by minimizing unfairness by a constrained re-ranking method that adds constraints over evaluation metrics. [33] argues that those post-processing methods that improve the diversity among recommendations contribute unfairness among users. They establish a method for measuring user satisfaction and employ the Gini coefficient of satisfaction scores to determine user unfairness. In a study by Fu et al. [17], a strategy is proposed that integrates fairness constraints with heuristic re-ranking methods. This approach aims to mitigate issues related to user unfairness in explainable recommendation systems that utilize knowledge graphs. [47] improves the user and item fairness by adding dummy data into the original user-item matrix. [55] presents a method for formulating the visibility of items and outlines a computational framework utilizing linear programming to optimize result ranking based on three forms of fairness constraints. In [71], Yang and Stoyanovich reduce variations in the distribution of item visibility across different groups using regularization techniques. Morik et al. [40] present a dynamic learning to a rank algorithm that addresses the rich-get-richer dynamics and allows for a configurable allocation of exposure scheme. [43] presents a post-processing method that re-ranks a given ranking to ensure that each item is guaranteed a minimum level of exposure opportunity. [34] proposes a GAN-based fairness-aware recommendation based on implicit feedback, which simultaneously solves the negative sampling issue and the item fairness issue. While these works have significantly contributed to solving various fairness scenarios in recommendations, a critical perspective has been largely overlooked: identifying the underlying causes of this unfairness. Understanding these causes is crucial for diagnosing the fairness issues inherent to the system and devising effective strategies for improvement. Hence, our study aims to bridge this gap by delving into the causes of unfairness in recommendation systems.

2.2 Explainable Recommendations

The goal of explainable recommendation is to create models that produce not only accurate recommendations, but also provide explanations that are easily understood [2, 6, 10, 38, 76]. Explainable recommendation systems can be classified into two categories: those that are intrinsic to the model, and those that are agnostic to the model being used [36, 39]. The model-intrinsic approach involves creating models that are interpretable and have a decision-making process that is clear, allowing for natural explanations of the model's decisions [77]. The model-agnostic approach, also known as the post-hoc explanation approach, enables the decision-making process to be a blackbox [44, 65]. For instance, the Explicit Factor Model for explainable recommendation [77] used a matrix factorization method [32] to generate recommendations and included an explanation sentence for each recommended item. On the other hand, the Interpretable Convolutional Neural Network approach [53] created a deep convolutional neural network model and showed item features to users as explanations. Another example is the visually explainable recommendation [13] that uses a deep model to create image regional-of-interest explanations. Recently, there have been proposals to use counterfactual reasoning to enhance explainability in recommendation systems [21, 59–61, 70]. As an illustration, Ghazimatin et al. [21] endeavor to generate provider-side counterfactual explanations by identifying the minimal set of a user's historical actions, such as

[1] https://github.com/jasonshere/ACFR

reviewing, purchasing, or rating, that, if removed, would lead to an altered recommendation. In [60], Tan et al. proposed generating and evaluating explanations based on causal relations to the outcome. Although these works have substantially advanced the field of explainable recommendations, they primarily focus on explaining the mechanisms by which recommendations are generated. However, a significant gap remains: the need to explain fairness in recommendations. This aspect is crucial because understanding the fairness of recommendations is equally as important as understanding the generation of these recommendations. The current work attempts to address this oversight by focusing on explaining the fairness of recommendations.

2.3 Fairness Explanations

Explainability and fairness are two crucial aspects for accountable recommendation systems, however, the relationship between the two is not yet fully understood. There have been several early studies attempting to provide explanations for fairness [9, 17, 18, 42]. As an example, the research conducted by Begley et al. [9] generates explanations by applying Shapley value paradigm [54] to identify the difference of the feature contributions to model disparity. Pan et al. [42] consider the causal relationships among feature variables and propose a novel framework to decompose the disparity into the sum of contributions from fairness-aware causal paths linking the sensitive attribute and the final predictions, on the graph. In [17], a fairness-conscious algorithm is proposed that utilizes a knowledge graph to provide transparent diversity in recommendations and possibly identify relevant items. Recently, [18] introduced a framework for explaining fairness in feature-based recommendations, utilizing a counterfactual reasoning paradigm.

Previous works on fairness explanations have strived to elucidate the root causes of unfairness. However, there's a glaring absence of literature that provides fairness explanations on an individual user and item level. Here's why this is crucial:

- **Foundation of Recommender Systems:** Interactions between users and items serve as the ground truth in recommendation model training [32, 49]. These interactions are indispensable, as they provide an accurate representation of user preferences and item relevancy, ensuring the efficacy of the recommendation model [5, 26].
- **Challenges with Feature-based Explanations:** While existing works have primarily focused on explaining fairness from features, two significant issues arise:
  - **Noise in Features:** Features used for training recommendation models might contain noise [77], which can lead to sub-optimal [8] recommendations if not handled properly. Relying solely on feature-based explanations can thus mislead or oversimplify the fairness landscape [18].
  - **Absence and Incompleteness:** Features can often be absent, incomplete, or inconsistent across different items or users, making them unreliable for fair explanations in some scenarios [4, 11, 52].

  By emphasizing interactions, our research circumvents these challenges, ensuring more precise and comprehensive fairness explanations. Given that interactions between individual users and items form the bedrock of recommender systems, understanding fairness at this granular level is paramount. Our research aims to fill this void by offering fairness insights tailored to individual users and items based on their unique interactions.

In summary, while existing literature provides valuable insights into fairness and explainability, our research offers a more nuanced approach, targeting the identified gaps and enhancing the overall understanding of fairness in recommender systems.

Symbols	Description
$\mathbf{G}$	The recommendation model for generating explanations
$\mathcal{G}$	The recommendation model for imputing missing interactions
$\mathcal{R}$	The sparse matrix of observed user-item interactions
$\hat{\mathcal{R}}$	The dense matrix of predicted interactions
$\hat{\mathcal{R}}_{\mathcal{G}}$	The predictions matrix generated by $\mathcal{G}$
$\tilde{\mathcal{R}}$	The dense matrix of imputed interactions
$\hat{\tilde{\mathcal{R}}}$	The predictions generated by $\mathbf{G}$ training on $\tilde{\mathcal{R}}$
$\hat{\tilde{\mathcal{R}}}^{cf}$	The counterfactual predictions generated by $\mathbf{G}$ after training on $\tilde{\mathcal{R}}$
$\mathcal{M}$	The weight matrix representing if an interaction is observed
e_i	The mean squared error over known interactions of user i
μ_i	The mean of all users' mean squared error
$\tilde{\mathbf{r}}$	The imputed interactions for a specific user/item
$\hat{\tilde{\mathbf{r}}}$	The predicted interactions generated by $\mathbf{G}$ on imputed interactions for a specific user/item
n	The number of users
m	The number of items
Ψ_{user}	The individual user fairness measure
Ψ_{item}	The individual item fairness measure
Ψ	The fairness measure based on the predictions generated by the pre-trained $\mathbf{G}$
Ψ^{cf}	The fairness measure based on the counterfactual predictions
ω	The weights of pre-trained $\mathbf{G}$
$\tilde{\omega}$	The weights of pre-trained $\mathbf{G}$ after retraining on imputaed interactions
$\Delta_{\tilde{\omega}}$	The changes of the weights of $\mathbf{G}$
$\tilde{\omega}^{cf}$	The counterfactual weights of $\mathbf{G}$

Table 1. Notations.

3 PRELIMINARY

In this work, we consider explaining both individual user unfairness and individual item unfairness in recommendations [2]. The notations used are explained in table 1.

3.1 Individual User Fairness

Individual user fairness encompasses the idea that each user should receive recommendations that are fair, unbiased, and have the same prediction accuracy. The individual user unfairness can be represented as the variance among all users' loss [47]:

$$\Psi_{user}(\mathcal{R}, \hat{\mathcal{R}}) = \frac{1}{n}\sum_{i=1}^{n}(e_i - \mu_e)^2, \tag{1}$$

where $\mathcal{R}$ is the user-item interaction matrix with rows corresponding to users and the columns corresponding to items, and the elements of $\mathcal{R}$ represent the corresponding rating or feedback given by the user to the item. $\mathcal{R}$ is sparse because most of the entries are empty, which means that most of the users have not rated most of the

[2]Both fairness and unfairness are used in different places of the paper for a better description. Generally, an increase in fairness means a decrease in unfairness and vice versa.

items; $\hat{\mathcal{R}}$ is a dense matrix with the same shape as $\mathcal{R}$ and each element corresponds to the predicted rating for a specific user-item pair; μ_e is the mean of all users' loss, n is the number of users, and e_i is the mean squared error over known ratings of user i, which is denoted as:

$$e_i = \frac{||\mathbf{r}_i - \hat{\mathbf{r}}_i||_2^2}{m_i}, \tag{2}$$

where m_i is the number of known ratings of user i.

3.2 Individual Item Fairness

Within our framework, individual item fairness is approached from the perspective of prediction parity, where the fairness metric is concerned with the equality of predicted ratings across items for each user [34, 40, 47]. This contrasts with an exposure-based fairness viewpoint, which considers the visibility and ranking of items in recommendation lists [15, 46, 51]. Given the predicted rating matrix $\hat{\mathcal{R}}$, we quantify individual item unfairness through the variance in predicted ratings for all items across users, as follows:

$$\Psi_{item}(\hat{\mathcal{R}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{j=1}^{m}(\hat{r}_{ij} - \mu_{\hat{r}_i})^2, \tag{3}$$

where $\mu_{\hat{r}_i}$ is the mean of predicted ratings of user i, m is the number of items. This metric thus reflects prediction parity fairness by evaluating the consistency of item ratings predictions made for each user, distinct from item exposure fairness.

Establishing prediction parity fairness as a fundamental concept in our work is pivotal. It serves as the groundwork from which we aim to extend into more complex analyses of exposure-based fairness. This progression underlines our belief that a solid understanding of prediction parity provides the necessary basis for delving into the nuances of how items are ranked and exposed to users, marking a critical path for our future research efforts in enhancing fairness within recommender systems.

User fairness emphasizes achieving parity in prediction accuracy across users, aiming to ensure equitable satisfaction among them by minimizing disparities in how accurately their preferences are predicted. This approach seeks to harmonize users' satisfaction levels, contributing to user fairness. Conversely, item fairness centers on the parity of predicted ratings across items, reflecting each item's opportunity to be recommended. Higher predicted ratings result in higher rankings, thereby increasing an item's visibility. Therefore, reducing prediction disparities across items levels the playing field, ensuring each has a fair chance of being recommended and achieving item fairness. Reducing item prediction parity should logically contribute to reducing item exposure-based unfairness, as items with more uniform prediction ratings would have a more balanced distribution in the recommendation lists. This intuition suggests that efforts to equalize the predicted ratings across items can indirectly enhance the fairness of item exposure, addressing one of the key mechanisms through which recommender systems influence item visibility and user engagement.

3.3 Justification and Contextualization of User and Item Fairness Metrics

The importance of fairness metrics in recommender systems lies in their ability to address different aspects of fairness: user fairness and item fairness. User fairness focuses on achieving parity in prediction accuracy across users, ensuring that each user receives recommendations tailored to their preferences without algorithmic bias [31, 35, 63]. This concept is grounded in the principle that equitable treatment in prediction accuracy leads to a higher degree of user satisfaction, aligning with the ultimate goals of personalization and user engagement [40].

Conversely, item fairness centers on prediction parity across items, aiming to give each item an equal opportunity to be recommended. While the objective of reducing item prediction disparity might appear to contradict the

purpose of personalized recommendations, the goal is more sophisticated. It involves ensuring that no item is systematically underrepresented or overrepresented due to biases inherent in the recommendation algorithm [7, 47]. This metric acts as a safeguard against biases that might skew the system's outputs, ensuring that the variations in item recommendation are justifiable by true user preferences rather than arbitrary model discrepancies [34].

To contextualize these metrics effectively within the ecosystem of recommender system evaluations, they should be considered alongside other key performance indicators such as accuracy, diversity, novelty, and overall user satisfaction [15, 46]. This balanced approach helps to maintain the effectiveness of the recommender system while ensuring fairness. User fairness metrics address the disparity in how the system serves different users, while item fairness metrics ensure an equal chance of recommendation for all items within the catalog. Integrating these considerations allows for a more comprehensive assessment of both personalization efficacy and fairness, providing a holistic view of system performance [51].

4 ADDING-BASED COUNTERFACTUAL FAIRNESS REASONING

To illuminate the relationship between fairness and users/items through their current interactions, we introduce an innovative solution named Adding-based Counterfactual Fairness Reasoning (*ACFR*). The *ACFR* process, detailed in algorithm 1, unfolds in three pivotal steps designed to systematically evaluate and enhance fairness in recommender systems:

- **Unknown Interaction Estimation and Imputation:** This initial step involves predicting potential interactions between users and items that are not present in the current dataset but could reasonably occur. These estimated interactions are then imputed into the dataset, enriching the existing user-item matrix with a more comprehensive view of potential engagements.
- **Adding-based Counterfactual Reasoning:** Building on the enriched dataset, this phase employs counterfactual analysis to explore hypothetical scenarios where certain user-item interactions are added. This step assesses how these additions would affect the fairness of the recommendation outcomes, allowing us to understand the specific contributions of different users and items to the overall fairness of the system.
- **Explanation Score Generation:** The final step quantifies the impact of each user and item on the system's fairness through explanation scores. These scores provide a numerical representation of the fairness contribution, distinguishing between entities that positively or negatively influence fairness. This metric facilitates a clear, actionable understanding of where interventions may be needed to promote fairness within the recommender system.

The *ACFR* methodology represents a significant advancement in fairness evaluation, offering a structured approach to dissecting the nuanced relationship between user-item interactions and fairness outcomes. By systematically implementing these steps, *ACFR* provides a detailed framework for identifying fairness imbalances and devising strategies to address them, enhancing the equity of recommendations provided by the system.

Central to *ACFR* is the intuition that fairness in recommender systems can be deeply influenced by the subtle dynamics of user-item interactions. By adopting a counterfactual approach, *ACFR* allows us to explore 'what-if' scenarios, shedding light on how changes in these interactions might affect fairness outcomes. This methodology underscores our belief that understanding and mitigating unfairness requires a detailed analysis of both the individual and collective impact of these interactions on the recommendation process.

4.1 Unknown Interaction Estimation and Imputation

Initially, we employ a two-step modeling process where our primary recommendation model, referred to as $\mathbf{G}$, is meticulously trained to capture the nuances of existing user-item interactions within our dataset. Parallel to this, we introduce a secondary model, also designated as $\mathcal{G}$ but specialized for imputation purposes, which undergoes

Algorithm 1 The process of *ACFR*

Input: The user-item ratings matrix $\mathcal{R}$, the number of steps $S_1 - step$, the number of steps $S_2 - step$.
Output: Explainability scores (ES) List $\mathbb{L}$
Train the recommendation model **G** by minimizing eq. (4).
Train the recommendation model $\mathcal{G}$ by minimizing eq. (4).
Calculate Ψ based on predictions produced by **G**
Save weights ω of **G**
for each user/item **do**
  Reset weights of **G** to ω
  Predict missing interactions of the user/item using $\mathcal{G}$
  Impute missing interactions for the user/item, then we obtain $\tilde{\mathbf{r}}$
  for $S_1 - step$ **do**
    Retrain **R** with $\tilde{\mathbf{r}}$ by minimizing eq. (7)
  end for
  Save $\tilde{\omega}$ of **G**
  for $S_2 - step$ **do**
    Find $\Delta_{\tilde{\omega}}$ by minimizing eq. (8)
  end for
  Calculate ES for the user/item via eq. (11)
  $\mathbb{L}$.append(**ES**)
end for

return $\mathbb{L}$

a similar training regimen on the same set of known interactions [48, 69]. The purpose of this sequential training approach is to leverage the predictive capabilities of the imputation model $\mathcal{G}$. It utilizes the learned patterns and relationships from the known interactions to accurately forecast the unknown or missing interactions across the user-item matrix. This predictive mechanism enables us to systematically impute these previously unidentified interactions, thus enriching the interaction matrix with a comprehensive set of user-item relationships for subsequent analysis. In this study, we utilize SVD++ [30, 64] (a classic matrix factorization model [32]) as **G**, and DeepFM [22] (a classic deep-learning-based recommendation model) as $\mathcal{G}$. Note that other recommendation models can be used as substitutes for **G** and $\mathcal{G}$.

We first train **G** and $\mathcal{G}$ on current user-item interactions by minimizing the mean squared error (MSE):

$$\mathcal{L}_{MSE} = ||\mathcal{R} - \mathcal{M} * \hat{\mathcal{R}}||_2^2, \tag{4}$$

where $\mathcal{R}$ is the user-item interaction matrix; $\hat{\mathcal{R}}$ is the prediction matrix; the matrix $\mathcal{M}$ is a weight matrix with the same shape as $\mathcal{R}$ and $\hat{\mathcal{R}}$, where an element is 1 if the corresponding user has rated the corresponding item, and 0 otherwise:

$$\mathcal{M}_i^j = \begin{cases} 1, & \text{if user } i \text{ interacted with item } j, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Given a particular user,[3] the vector after imputing interactions is denoted as $\tilde{\mathbf{r}} \in \tilde{\mathcal{R}}$:

$$\tilde{\mathcal{R}} = \mathcal{M} * \mathcal{R} + (1 - \mathcal{M}) * \hat{\mathcal{R}}_{\mathcal{G}}, \tag{6}$$

[3]The process for an item is same as that for a user unless specified otherwise. For the sake of description conciseness, we discuss user only in many places in this Section.

where $\hat{\mathcal{R}}_{\mathcal{G}}$ is the predictions matrix generated by pre-trained $\mathcal{G}$. Eq. 6 outlines the process of maintaining the original, known ratings within the matrix while substituting the missing ratings with predictions from $\mathcal{G}$, thereby enabling a comprehensive evaluation that integrates both actual user-item interactions and predicted preferences. This approach ensures a balanced assessment, reflecting the system's performance across known interactions and potential recommendations, crucial for a nuanced understanding of fairness implications.

4.2 Adding-based Counterfactual Reasoning

The adding-based counterfactual reasoning aims to examine the impact of imputed interactions at the level of individual users to recommendation fairness. Imputing interactions for some users may lead to the positive impact of recommendation fairness while that for others may lead to negative impact.

Suppose we have imputed interactions for a particular user as eq. (6). The original and imputed interactions of the user are represented by $\tilde{\mathbf{r}} \in \tilde{\mathcal{R}}$. Then, by treating $\tilde{\mathbf{r}}$ as the real interactions of the user, we can update the weights of the pre-trained model $\mathbf{G}$ from ω to $\tilde{\omega}$. This can be done by minimizing the mean squared error between the predicted interactions of the user using $\mathbf{G}$, denoted as $\hat{\tilde{\mathbf{r}}}$, and the interactions in $\tilde{\mathbf{r}}$:

$$\mathcal{L}(\hat{\tilde{\mathbf{r}}}) = ||\tilde{\mathbf{r}} - \hat{\tilde{\mathbf{r}}}||_2^2, \tag{7}$$

Then we slightly intervene the weights $\tilde{\omega}$ of $\mathbf{G}$ with $\Delta_{\tilde{\omega}}$. That is, the weights of $\mathbf{G}$ turns to $\tilde{\omega}^{cf}$ by adding $\Delta_{\tilde{\omega}}$ to the weights $\tilde{\omega}$ of $\mathbf{G}$. It is worth mentioning that the reason we consider intervening all weights of $\mathbf{G}$ is that retraining $\mathbf{G}$ with $\tilde{\mathbf{r}}$ will affect all weights of $\mathbf{G}$, not just a subset of them. With the new weights $\tilde{\omega}^{cf}$, $\mathbf{G}$ will change the predictions from $\hat{\tilde{\mathcal{R}}}$ to the counterfactual predictions $\hat{\tilde{\mathcal{R}}}^{cf}$. The counterfactual predictions will change the fairness measure from Ψ to Ψ^{cf}. The objective of adding-based counterfactual reasoning is to determine the smallest adjustments to the weights $\tilde{\omega}$ that will lead to the greatest decrease in Ψ, by minimizing the objective function:

$$\min \Psi^{cf} + \alpha \cdot ||\Delta_{\tilde{\omega}}||_2^2, \tag{8}$$

where α is a hyper-parameter to control the weight between the two terms. The first term of the objective function is intended to achieve the greatest decrease in unfairness, while the second term represents the difference between the original weights and the counterfactual weights. It is worth mentioning that identifying changes in interactions $\tilde{\mathbf{r}}$ directly can be difficult as it is challenging to pass gradients directly to $\tilde{\mathbf{r}}$ when minimizing Ψ^{cf}.

If the goal is to generate explanations for user fairness as defined in eq. (1), the fairness measure Ψ is denoted as Ψ_{user} and Ψ^{cf} is denoted as Ψ^{cf}_{user}. If the goal is to generate explanations for item fairness as defined in eq. (3), Ψ is denoted as Ψ_{item} and Ψ^{cf} is denoted as Ψ^{cf}_{item}.

4.3 Explainability Score Generation

For each user, after solving the counterfactual optimization problem by minimizing eq. (8), the minimal change $\Delta_{\tilde{\omega}}$ and the predictions of the corresponding interactions based on the adjusted weights of $\mathbf{G}$ will be obtained. Following [18], Validity $\mathcal{V}$ is calculated to measure the impact on recommendation fairness caused by the weights' perturbation:

$$\mathcal{V} = \Psi - \Psi^{cf}, \tag{9}$$

and Proximity $\mathcal{P}$ is then calculated to measure the degree of change in the weights of $\mathbf{G}$, which is represented by the difference between the weights updated on $\tilde{\mathbf{r}}$ and the counterfactual weights:

$$\mathcal{P} = ||\Delta_{\tilde{\omega}}||_2^2. \tag{10}$$

Then the explainability score (ES) can be calculated for each user or item based on their corresponding counterfactuals:

$$ES = \mathcal{V} - \beta \cdot \mathcal{P}, \tag{11}$$

where $\beta \in (0, 1)$ is a hyper-parameter to control the importance of Validity. The ES measures how effective a user/item's interactions are at reducing unfairness in model **G**. A higher ES indicates a greater ability to improve fairness.

5 EMPIRICAL EVALUATION

5.1 Datasets

To evaluate the effectiveness of the proposed *ACFR*, we followed [18, 33, 34] and conducted experiments on three real-world datasets from Amazon[4] [25] and MovieLens[5] [24]:

- *Toys and Games*, which includes 2,252,771 interactions from 1,342,911 users on 327,698 items;
- *Beauty*, which contains 2,023,070 interactions between 1,210,271 users and 249,274 items;
- *MovieLens 100k*, which contains 100,000 interactions between 943 users and 1682 items.
- *MovieLens 25M*, which contains 25 million interactions between 62,000 movies by 162,000 users, released in December of 2019.

Similar to MovieLens [24], we remove users and items with less than 10 interactions for Toys and Games, and Beauty datasets. For the experiments, we used a 5-fold cross-validation method and report the average results on the test dataset.

5.2 Baselines

The core application of *ACFR* is identifying the key users and items which are related to recommendation fairness, either user fairness or item fairness. Since no method exists to solve the same problem, we adopt the following methods as baselines:

- **Random-User** selects N users from all users at random, without replacement.
- **Random-Item** selects N items from all items at random, without replacement.
- **Active-User** selects the top N most active users from the entire set of users.
- **Inactive-User** selects the top N least active users from the entire set of users.
- **Active-Item** selects the top N most active items from the entire set of items.
- **Inactive-Item** selects the top N least active items from the entire set of items.

For the proposed *ACFR*, we set the hyper-parameter $\alpha = 0.01$ (eq. (8)) and $\beta = 0.07$ (eq. (11)). The model parameters of **G** and $\mathcal{G}$ are optimized by Adam optimizer. The learning rate for training **G** is 1e-4, and the learning rate for training $\mathcal{G}$ is 1e-3. For both **G** and $\mathcal{G}$, the dimension of latent factors is set as 20. The number of hidden units in each layer of DNN of $\mathcal{G}$ is set to 16, and we set the number of layers as 2. We use $relu$ as the activation function of DNN layers. For the regularization and bias coefficients specific to **G**, we adopted a value of 1e-3, aimed at preventing overfitting while allowing the model sufficient flexibility to capture the underlying patterns in the data. Additionally, a dropout rate of 0.1 was applied to $\mathcal{G}$ to further mitigate the risk of overfitting by randomly omitting a portion of the feature detectors on each training pass.

In our investigation of fairness within recommender systems, we introduce eight distinct variants of the *ACFR* framework. Each variant is tailored to assess and elucidate fairness from the unique perspectives of either users or items, employing a nuanced analysis that accounts for their individual contributions to fairness or unfairness. The designation of "*User*" or "*Item*" following "*ACFR-*" explicitly signifies the focus of the analysis on either users or items. The subscripts u and i within *ACFR-*$^*_u$ or *ACFR-*$^*_i$ categorically indicate the target of the fairness evaluation—user fairness or item fairness, respectively. Additionally, the symbols "(-)" and "(+)" are used to

[4]https://cseweb.ucsd.edu/ jmcauley/datasets.html
[5]https://grouplens.org/datasets/movielens/

distinguish between the entities considered to have a negative or positive impact on fairness, respectively, based on their Explanation Score (ES):

- *ACFR* - User_u(-): This variant is designed to identify a subset of users deemed to have the most significant negative impact on user fairness. It focuses on those with the lowest ES, suggesting they contribute less to promoting fairness within the system.
- *ACFR* - User_u(+): Conversely, this variant targets users who are assessed to positively influence user fairness the most, indicated by having the highest ES. These users are considered pivotal in enhancing the fairness of the recommender system.
- *ACFR* - User_i(-): Similar to the first variant, but with a focus on item fairness, this variant identifies users who negatively affect item fairness, again based on the lowest ES.
- *ACFR* - User_i(+): This variant identifies users who have a positive impact on item fairness, selecting those with the highest ES among the user base, suggesting their interactions promote a more equitable item representation.
- *ACFR* - Item_u(-): Targets items that, based on their ES, are determined to have the most substantial negative impact on user fairness. These items are identified as contributing least towards equitable user experiences.
- *ACFR* - Item_u(+): Focuses on items that positively affect user fairness, selecting those with the highest ES. Such items are crucial for ensuring users receive fair and balanced recommendations.
- *ACFR* - Item_i(-): Identifies items with the lowest ES that negatively impact item fairness, highlighting items that might contribute to an imbalance in item visibility or representation.
- *ACFR* - Item_i(+): Identifies items with the highest ES that positively influence item fairness, ensuring a diverse and equitable distribution of item recommendations across the user base.

This structured approach allows for a granular analysis of fairness within recommender systems, illuminating the roles individual users and items play in shaping the fairness landscape. By quantifying the impact of specific entities through their Explanation Scores (ES), the *ACFR* framework offers a nuanced tool for understanding and addressing fairness in recommendations.

5.3 Evaluation Methods

The erasing-based counterfactual analysis method in the previous work [18] measures how much the recommendation performance would change after the "most important" features are removed. Differently, we use an adding-based counterfactual analysis method that measures how the recommendation fairness changes after adding additional interactions to the selected N users/items. For all baselines and all variants of *ACFR*, the unknown interactions of the selected N users/items are predicted and imputed using model $\mathcal{G}$, then the adding-based counterfactual analysis is based on training the recommendation model **G** before and after maximizing recommendation fairness as discussed in section 4. In this study, we chose to use N=10, N=20, and N=30.

5.4 Metrics

To evaluate the quality of our recommendations, we rely on three commonly used metrics: **Precision** [45], Normalized Discounted Cumulative Gain (**NDCG**) [16, 29], and Root Mean Square Error (**RMSE**) [28]. Precision measures the proportion of recommended items that are relevant, NDCG evaluates the effectiveness of the ranking of recommended items by taking into account their relevance and position, and RMSE measures the accuracy of the recommendation system by comparing predicted ratings with actual ratings. We measure the performance of Precision and NDCG at a recommendation length of **5**. A higher value of Precision and NDCG indicates better recommendation quality. For RMSE, a smaller value represents the better recommendations accuracy. Additionally, we measure the fairness of our recommendations using two metrics: one that assesses

unfairness among individual users (using eq. (1)) and another that evaluates individual item unfairness (using eq. (3)). A smaller value for both user and item unfairness metrics represents better recommendation fairness.

5.5 Experimental Results

In this section, we provide the results of the evaluation metrics - Precision@5, NDCG@5, RMSE, User Unfairness, and Item Unfairness - that were obtained from the predictions of $\mathbf{G}$. The model was retrained, incorporating the imputed interactions of selected users or items (with N=10, 20, and 30). We present the outcomes separately from the perspectives of users and items. Tables table 2 and table 3 gauge the ability of *ACFR* to discern users and items that have a beneficial impact on both user fairness and item fairness.

Conversely, Tables table 4 and table 5 test the potential of *ACFR* to detect users and items that impart a detrimental effect on both user fairness and item fairness. To ensure a fair comparison, we juxtapose the user-focused variants of *ACFR* with the user-focused baselines (i.e., Random-User, Active-User, and Inactive-User). Similarly, we put the item-focused variants of *ACFR* up against the item-focused baselines (i.e., Random-Item, Active-Item, and Inactive-Item).

5.5.1 Analysis of Users Positively Impacting Fairness. In this section, we aim to elucidate the adeptness of our proposed system, *ACFR*, in identifying users who make a positive contribution to both user and item fairness. We turn to the results presented in table 2 to provide a concrete illustration of this capability. In our analysis focusing on user fairness across different settings, we find that the Inactive-User baseline outperforms other baseline methods at N=10, 20, and 30, particularly within the *Toys and Games* and *Beauty* datasets. This observation points to the Inactive-User baseline's effective performance in promoting user fairness under these specific evaluation settings. However, when we turn our attention to the *MovieLens 100K* dataset, the picture becomes less clear-cut. In this case, no single baseline model demonstrates a significantly superior performance in user fairness than the others, which highlights the complexities and diverse user behaviors inherent in different datasets. When it comes to item fairness, the Active-User leads the pack on the *Beauty* and *MovieLens 100k* datasets. On the *Toys and Games* dataset, however, the landscape shifts and we observe both the Random-User and Inactive-User outperforming the rest.

In the face of stiff competition, the *ACFR* variants, denoted as *ACFR* - $\text{User}_u(+)$ and *ACFR* - $\text{User}_i(+)$, stand out. The former excels across all datasets in terms of user fairness, while the latter surpasses all baselines in item fairness on every dataset examined. This exemplary performance solidly establishes the prowess of *ACFR* in accurately pinpointing users who have a positive bearing on either user or item fairness. The demonstrated superiority of *ACFR* reaffirms its value in explaining fairness in recommender systems. By effectively detecting users contributing positively to the overall fairness, *ACFR* presents a concrete step forward in balancing the often competing demands of personalized recommendation and fairness.

It is imperative to note, however, that the primary goal of our investigation is not to directly enhance fairness metrics but to illuminate the roles individual users and items play in affecting fairness—both positively and negatively—within the ecosystem of a recommender system. Observations related to changes in accuracy metrics, as a result of prioritizing fairness, fall outside the direct purview of our study. Our focus lies in identifying entities that contribute to or detract from fairness, thereby offering nuanced insights into the dynamics of fairness in recommender systems. This analytical perspective allows us to delve deeper into the mechanics of recommendation fairness, underscoring our commitment to advancing understanding in this critical area of research.

5.5.2 Analysis of Items Positively Impacting Fairness. In this section, we broaden our investigation to determine how well *ACFR* performs in identifying items that positively influence both user and item fairness. This element is pivotal to enhancing the overall fairness of the recommendation system. The results exhibited in table 3 indicate

Methods	Precision@5 (%)			NDCG@5 (%)			RMSE			User Unfairness			Item Unfairness		
	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30
Toys and Games															
The Model **G**	0.316			0.372			0.959			1.280			0.136		
Random-User	0.339	0.361	0.293	0.391	0.405	0.350	0.943	0.939	0.940	1.191	1.171	1.169	0.135	0.135	0.135
Active-User	**0.384**	**0.361**	**0.361**	**0.426**	**0.416**	**0.395**	0.951	0.948	0.945	1.254	1.238	1.235	0.138	0.137	0.137
Inactive-User	0.316	0.339	0.271	0.359	0.390	0.332	0.941	**0.938**	**0.937**	1.179	1.157	1.153	0.136	0.134	0.135
ACFR - User$_u$(+)	0.316	0.316	0.316	0.389	0.389	0.387	**0.939**	0.939	0.939	**1.149**	**1.148**	**1.147**	0.134	0.136	0.136
ACFR - User$_i$(+)	0.339	0.316	0.316	0.386	0.375	0.375	0.942	0.940	0.940	1.207	1.193	1.196	**0.132**	**0.132**	**0.132**
Beauty															
The Model **G**	0.108			0.143			0.895			1.212			0.178		
Random-User	**0.171**	0.109	0.140	**0.264**	0.167	0.211	0.889	0.889	**0.887**	1.181	1.134	1.134	0.164	0.171	0.166
Active-User	0.155	**0.171**	**0.202**	0.256	**0.243**	**0.263**	0.892	0.890	0.890	1.216	1.214	1.210	0.161	0.162	0.162
Inactive-User	0.155	0.155	0.124	0.221	0.218	0.160	**0.889**	**0.888**	0.888	1.145	1.126	1.115	0.173	0.172	0.174
ACFR - User$_u$(+)	0.109	0.109	0.109	0.144	0.153	0.153	0.890	0.890	0.890	**1.112**	**1.100**	**1.097**	0.177	0.177	0.177
ACFR - User$_i$(+)	0.140	0.140	0.124	0.193	0.212	0.182	0.891	0.889	0.888	1.214	1.204	1.204	**0.153**	**0.153**	**0.153**
MovieLens 100k															
The Model **G**	2.632			2.775			0.942			0.399			0.310		
Random-User	**3.992**	3.270	3.652	**4.170**	3.466	3.952	0.943	0.941	0.941	0.394	0.383	0.390	0.280	0.252	0.264
Active-User	3.907	3.524	3.482	4.095	3.702	3.688	**0.942**	**0.941**	**0.941**	0.396	0.388	0.387	0.260	0.253	0.252
Inactive-User	3.652	**3.949**	**4.076**	3.879	**4.113**	**4.195**	0.942	0.943	0.942	0.389	0.389	0.388	0.298	0.294	0.285
ACFR - User$_u$(+)	2.696	2.696	2.718	2.839	2.838	2.887	0.946	0.944	0.944	**0.374**	**0.374**	**0.373**	0.257	0.243	0.246
ACFR - User$_i$(+)	2.803	2.845	2.718	2.943	2.974	2.857	0.945	0.945	0.945	0.383	0.384	0.382	**0.229**	**0.228**	**0.228**

Table 2. Analysis of Users Positively Impacting Fairness: Accuracy (Precision@5, NDCG@5, RMSE) and Fairness (User Unfairness, Item Unfairness) - Best Results Highlighted in Bold.

that, concerning user fairness, the Random-Item stands out by surpassing other baselines on the *MovieLens 100k* datasets. On the *Toys and Games* dataset, Random-Item outperforms other baselines on user fairness when N equals 10 and 20. However, when N is 30, the Inactive-Item takes the lead in terms of user fairness. Similarly, on the *Beauty* dataset, the Inactive-Item outperforms all other baselines. When it comes to item fairness, the Active-Item consistently outperforms other baselines across different values of N (10, 20, and 30) on the *Toys and Games* dataset, and for N values of 20 and 30 on the *Beauty* dataset. As for the *MovieLens 100k* dataset, the Random-Item surpasses all other baselines.

What stands as an endorsement of *ACFR*'s robustness is the performance of its variants, *ACFR* - Item$_u$(+) and *ACFR* - Item$_i$(+). Both these variants outperform all the baselines in terms of user fairness and item fairness across all the datasets. This solidifies the fact that *ACFR* is proficient at identifying those items that have a positive influence on both aspects of fairness, thus leading to more equitable recommendations.

Methods	Precision@5 (%)			NDCG@5 (%)			RMSE			User Unfairness			Item Unfairness		
	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30
Toys and Games															
The Model G	0.316			0.372			0.959			1.280			0.136		
Random-Item	**0.406**	**0.339**	0.316	**0.471**	**0.404**	0.378	0.942	0.939	0.939	1.166	1.151	1.152	0.140	0.141	0.140
Active-Item	0.316	0.271	0.293	0.416	0.359	0.377	0.946	0.943	0.941	1.200	1.180	1.176	0.137	0.137	0.138
Inactive-Item	0.361	0.271	0.316	0.404	0.356	0.396	0.943	0.939	0.937	1.187	1.162	1.146	0.139	0.140	0.141
ACFR - Item$_u$(+)	0.293	0.248	0.293	0.366	0.328	0.366	**0.937**	**0.937**	**0.937**	**1.139**	**1.137**	**1.134**	0.147	0.147	0.147
ACFR - Item$_i$(+)	0.361	0.293	**0.339**	0.405	0.351	**0.407**	0.958	0.958	0.956	1.283	1.283	1.272	**0.129**	**0.126**	**0.126**
Beauty															
The Model G	0.108			0.143			0.895			1.212			0.178		
Random-Item	0.155	0.124	0.109	0.181	0.169	0.137	0.889	0.886	0.885	1.131	1.083	1.072	0.173	0.184	0.188
Active-Item	0.124	0.124	0.124	0.162	0.168	0.161	**0.887**	**0.886**	**0.885**	1.106	1.098	1.094	0.182	0.180	0.180
Inactive-Item	0.124	0.140	0.140	0.186	0.204	0.206	0.888	0.887	0.887	1.054	1.055	1.049	0.199	0.199	0.198
ACFR - Item$_u$(+)	0.124	0.124	0.124	0.159	0.162	0.170	0.887	0.887	0.887	**1.050**	**1.046**	**1.048**	0.203	0.205	0.205
ACFR - Item$_i$(+)	**0.218**	**0.218**	**0.233**	**0.271**	**0.301**	**0.310**	0.895	0.894	0.893	1.224	1.200	1.200	**0.166**	**0.166**	**0.164**
MovieLens 100k															
The Model G	2.632			2.775			0.942			0.399			0.310		
Random-Item	2.611	2.718	2.611	2.720	2.861	2.683	**0.942**	**0.940**	**0.940**	0.385	0.384	0.384	0.294	0.298	0.301
Active-Item	2.357	2.357	2.378	2.429	2.400	2.464	0.943	0.942	0.942	0.408	0.406	0.406	0.307	0.306	0.304
Inactive-Item	**2.972**	**2.866**	**2.909**	**3.238**	**3.125**	**3.093**	0.942	0.942	0.942	0.391	0.385	0.385	0.314	0.327	0.331
ACFR - Item$_u$(+)	2.463	2.399	2.442	2.527	2.439	2.426	0.944	0.944	0.945	**0.383**	**0.379**	**0.377**	0.308	0.319	0.325
ACFR - Item$_i$(+)	2.654	2.633	2.633	2.804	2.790	2.790	0.949	0.949	0.949	0.419	0.421	0.421	**0.268**	**0.266**	**0.265**

Table 3. Analysis of Items Positively Impacting Fairness: Accuracy (Precision@5, NDCG@5, RMSE) and Fairness (User Unfairness, Item Unfairness) - Best Results Highlighted in Bold.

5.5.3 Analysis of Users Negatively Impacting Fairness. This section is dedicated to evaluating *ACFR*'s proficiency in pinpointing users that exert a negative impact on both user fairness and item fairness. As such, our attention is geared towards the methods that display the poorest performance concerning user fairness or item fairness. As per the results presented in table 4, the Active-User consistently lags behind the other baselines concerning user fairness on the *Toys and Games* and *Beauty* datasets. On the *MovieLens 100k* dataset, the Inactive-User fares the worst in terms of user fairness when N is set to 20 and 30, whereas the Active-User shows the poorest performance at $N = 10$. In the context of item fairness, the Active-User underperforms on the *Toys and Games* dataset, with the Inactive-User displaying the weakest performance on the *Beauty* and *MovieLens 100k* datasets.

The performance of *ACFR*'s variants, specifically *ACFR* - User$_u(-)$ and *ACFR* - User$_i(-)$ (values highlighted with an underline), provides an intriguing insight. These variants manifest the weakest performance among all baselines across all datasets concerning user fairness and item fairness, individually, even trailing behind

Methods	Precision@5 (%)			NDCG@5 (%)			RMSE			User Unfairness			Item Unfairness		
	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30
Toys and Games															
The Model **G**	0.316			0.372			0.959			1.280			0.136		
Random-User	0.339	0.361	0.293	0.391	0.405	0.350	0.943	0.939	0.940	1.191	1.171	1.169	**0.135**	0.135	0.135
Active-User	0.384	0.361	0.361	**0.426**	0.416	0.395	0.951	0.948	0.945	1.254	1.238	1.235	0.138	0.137	0.137
Inactive-User	0.316	0.339	0.271	0.359	0.390	0.332	**0.941**	**0.938**	**0.937**	**1.179**	**1.157**	**1.153**	0.136	**0.134**	**0.135**
ACFR - User$_u(-)$	**0.384**	0.384	0.384	0.449	**0.445**	0.442	0.956	0.955	0.954	1.269	1.275	1.272	0.137	0.138	0.138
ACFR - User$_i(-)$	0.339	**0.384**	**0.429**	0.421	0.420	**0.474**	0.950	0.948	0.949	1.244	1.226	1.235	0.142	0.142	0.141
Beauty															
The Model **G**	0.108			0.143			0.895			1.212			0.178		
Random-User	0.171	0.109	0.140	0.264	0.167	0.211	0.889	0.889	**0.887**	1.181	1.134	1.134	0.164	0.171	0.166
Active-User	0.155	0.171	**0.202**	0.256	0.243	0.263	0.892	0.890	0.890	1.216	1.214	1.210	0.161	0.162	0.162
Inactive-User	0.155	0.155	0.124	0.221	0.218	0.160	**0.889**	**0.888**	0.888	**1.145**	**1.126**	**1.115**	0.173	0.172	0.174
ACFR - User$_u(-)$	**0.202**	**0.186**	0.186	**0.270**	**0.273**	**0.291**	0.894	0.893	0.892	1.244	1.244	1.239	**0.159**	**0.158**	**0.157**
ACFR - User$_i(-)$	0.140	0.140	0.124	0.200	0.211	0.200	0.893	0.893	0.893	1.189	1.188	1.187	0.177	0.176	0.176
MovieLens 100k															
The Model **G**	2.632			2.775			0.942			0.399			0.310		
Random-User	**3.992**	3.270	3.652	**4.170**	3.466	3.952	0.943	0.941	0.941	0.394	**0.383**	0.390	0.280	**0.252**	0.264
Active-User	3.907	3.524	3.482	4.095	3.702	3.688	**0.942**	**0.941**	**0.941**	0.396	0.388	**0.387**	**0.260**	0.253	**0.252**
Inactive-User	3.652	3.949	**4.076**	3.879	4.113	4.195	0.942	0.943	0.942	**0.389**	0.389	0.388	0.298	0.294	0.285
ACFR - User$_u(-)$	3.928	**3.970**	4.013	4.056	**4.189**	**4.257**	0.943	0.942	0.942	0.403	0.398	0.396	0.266	0.262	0.261
ACFR - User$_i(-)$	3.843	3.779	3.694	4.067	4.035	3.864	0.943	0.942	0.942	0.400	0.395	0.395	0.319	0.313	0.316

Table 4. Analysis of Users Negatively Impacting Fairness: Results on accuracy (Precision@5, NDCG@5 and RMSE) and fairness (User Unfairness and Item Unfairness). The best results are bold-faced.

the performance of Model **G**. This underscores *ACFR*'s versatility in not only recognizing users who contribute positively to fairness but also identifying those who negatively influence fairness.

5.5.4 Analysis of Items Negatively Impacting Fairness. In this section, our focus shifts to examining *ACFR*'s capability in pinpointing items that create adverse effects on both user fairness and item fairness. Our emphasis, similar to the previous section, remains on the methods that exhibit subpar performance in terms of user fairness or item fairness. The results presented in table 5 indicate that the Active-Item typically falls short of other baselines concerning user fairness on the *Toys and Games* and *MovieLens 100k* datasets. On the *Beauty* dataset, the performance of the Random-Item is the least impressive in terms of user fairness when $N = 10$. However, it is the Active-Item that shows the poorest performance when $N = 20$ and $N = 30$. From the perspective of item fairness, the Inactive-Item lags on the *Beauty* and *MovieLens 100k* datasets. On the *Toys and Games* dataset, the performance of the Random-Item and Inactive-Item appears to be comparable.

Methods	Precision@5 (%)			NDCG@5 (%)			RMSE			User Unfairness			Item Unfairness		
	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30	N=10	N=20	N=30
Toys and Games															
The Model **G**	0.316			0.372			0.959			1.280			0.136		
Random-Item	**0.406**	**0.339**	0.316	**0.471**	**0.404**	0.378	0.942	0.939	0.939	1.166	1.151	1.152	0.140	0.141	0.140
Active-Item	0.316	0.271	0.293	0.416	0.359	0.377	0.946	0.943	0.941	1.200	1.180	1.176	0.137	0.137	0.138
Inactive-Item	0.361	0.271	0.316	0.404	0.356	**0.396**	0.943	0.939	0.937	1.187	1.162	1.146	0.139	0.140	0.141
ACFR - Item$_u(-)$	0.361	0.271	**0.339**	0.374	0.340	0.393	0.956	0.957	0.957	1.276	1.283	1.282	**0.132**	**0.130**	**0.129**
ACFR - Item$_i(-)$	0.293	0.293	0.293	0.366	0.366	0.366	**0.936**	**0.936**	**0.935**	**1.133**	**1.135**	**1.131**	0.147	0.147	0.148
Beauty															
The Model **G**	0.108			0.143			0.895			1.212			0.178		
Random-Item	0.155	0.124	0.109	0.181	0.169	0.137	0.889	0.886	0.885	1.131	1.083	1.072	0.173	0.184	0.188
Active-Item	0.124	0.124	0.124	0.162	0.168	0.161	**0.887**	**0.886**	**0.885**	1.106	1.098	1.094	0.182	0.180	0.180
Inactive-Item	0.124	0.140	0.140	0.186	0.204	0.206	0.888	0.887	0.887	1.054	1.055	1.049	0.199	0.199	0.198
ACFR - Item$_u(-)$	**0.155**	**0.155**	**0.155**	**0.202**	**0.220**	**0.237**	0.891	0.889	0.888	1.152	1.139	1.126	**0.168**	**0.166**	**0.168**
ACFR - Item$_i(-)$	0.124	0.124	0.124	0.170	0.170	0.170	0.887	0.888	0.887	**1.047**	**1.045**	**1.044**	0.205	0.206	0.206
MovieLens 100k															
The Model **G**	2.632			2.775			0.942			0.399			0.310		
Random-Item	2.611	2.718	2.611	2.720	2.861	2.683	**0.942**	**0.940**	**0.940**	0.385	0.384	0.384	0.294	0.298	0.301
Active-Item	2.357	2.357	2.378	2.429	2.400	2.464	0.943	0.942	0.942	0.408	0.406	0.406	0.307	0.306	0.304
Inactive-Item	**2.972**	**2.866**	**2.909**	**3.238**	**3.125**	**3.093**	0.942	0.942	0.942	0.391	0.385	0.385	0.314	0.327	0.331
ACFR - Item$_u(-)$	2.590	2.718	2.611	2.679	2.843	2.767	0.943	0.947	0.946	0.401	0.410	0.407	**0.294**	**0.270**	**0.272**
ACFR - Item$_i(-)$	2.505	2.484	2.484	2.574	2.569	2.551	0.946	0.946	0.946	**0.384**	**0.384**	**0.384**	0.354	0.354	0.354

Table 5. Analysis of Items Negatively Impacting Fairness: Results on accuracy (Precision@5, NDCG@5 and RMSE) and fairness (User Unfairness and Item Unfairness). The best results are bold-faced.

Upon examining the performance of the specific variants of *ACFR*, namely *ACFR* - Item$_u(-)$ and *ACFR* - Item$_i(-)$ (values highlighted with an underline), we discern a compelling trend. These variants uniformly register the least impressive performance among all baselines across every dataset, irrespective of whether we consider user fairness or item fairness, and even fall short of Model **G**′s performance. This underlines the multifaceted capacity of *ACFR* in not just identifying items that contribute positively to fairness, but also recognizing those that cast a negative influence on fairness.

A noteworthy observation is that *ACFR* - Item$_i(-)$ exhibits the best performance concerning user fairness on all datasets, while *ACFR* - Item$_u(-)$ shines in terms of item fairness across all datasets. This implies a peculiar dynamic: items that negatively influence user fairness seem to have a positive impact on item fairness, and vice versa. Items that cast a negative influence on item fairness appear to enhance user fairness.

5.5.5 Discussion on the Trends across N=10/20/30. Understanding the trends and variations observed at different levels of N, particularly at $N = 10$, $N = 20$, and $N = 30$, requires a closer look into the potential factors influencing these outcomes. Two primary reasons emerge:

- **Diverse Fairness Impact among Top Users/Items:** The set of N users or items likely contain a mixture, where some elements play a positive role in influencing fairness while others may contribute negatively.
- **Inherent Characteristics of *ACFR*'s Explainability Score (ES):** While *ACFR* demonstrates capability in capturing the correlation between ES and fairness, it may not always pinpoint an exact ES for every user or item. As illustrated in fig. 6b (where the x-axis represents ES values, the y-axis denotes user unfairness, and each dot symbolizes an individual user), there's a variance in how users with comparable ES influence fairness. However, the robust Pearson's Correlation of -0.905 between user unfairness and ES underlines its high negative correlation, emphasizing that despite some inconsistencies, the ES computed by *ACFR* remains intrinsically linked to fairness.

With these reasons as a backdrop, specific patterns in the data can be elucidated:

- **Observation from the Beauty Dataset (table 3):** *ACFR* - $\text{Item}_u(+)$ on the Beauty dataset demonstrates a pattern of user unfairness decreasing from $N = 10$ to $N = 20$, followed by an increase from $N = 20$ to $N = 30$. This aligns with the first reason. The top set of 20 items seem to have a pronounced capability to ameliorate user fairness. However, on expanding the scope to 30 items, certain newly integrated items may start exerting a negative influence on user fairness.
- **Insights from the Toys and Games Dataset (table 4):** *ACFR* - $\text{User}_u(-)$ shows a trend where user unfairness climbs from $N = 10$ to $N = 20$ and subsequently falls from $N = 20$ to $N = 30$. This trend can be ascribed to the first reason, indicating that all users from $N = 10$ to $N = 20$ predominantly hamper user fairness. Yet, a subset of users added between $N = 20$ to $N = 30$ might have started imparting a positive effect on user fairness.
- **Further Observations from Toys and Games & MovieLens 100k (table 5):** Noteworthy is the performance trend of *ACFR* - $\text{Item}_u(-)$ in relation to user unfairness. For both datasets, an increase from $N = 10$ to $N = 20$ is followed by a subsequent decrease from $N = 20$ to $N = 30$. Rooted in the first reason, this suggests that items within the $N = 10$ to $N = 20$ bracket potentially detract from user fairness. Conversely, some items in the $N = 20$ to $N = 30$ range seem to be positively influencing fairness.
- **Observations on the MovieLens 100k Dataset (table 4):** In the context of item unfairness, a distinctive pattern is observed with *ACFR* - $\text{User}_i(-)$. While the metric exhibits a decrease at $N = 20$, it subsequently increases at $N = 30$. Informed by the second reason, it's plausible that users within the $N = 10$ to 20 range comprise individuals who contribute positively to item fairness. This could counterbalance some of the adverse effects from the top 10 users. However, the rise observed at $N = 30$ suggests that the newly included users might exert a more potent negative influence on item fairness, diminishing the positive impacts from the earlier group.

5.5.6 Understanding the Interplay between Accuracy and Fairness. In this part of the discussion, we will navigate through scenarios where there is a mutual elevation in accuracy (Quantified through Precision@5, NDCG@5, RMSE) and fairness (user-focused and item-focused). As corroborated by preceding studies [20, 27, 73], augmenting additional information through these interactions potentially boosts the accuracy in the generation of recommendations. Our proposed system, *ACFR*, makes a significant contribution in this regard by successfully detecting those users or items for whom imputed interactions result in a positive change in user fairness or item fairness. However, we must make a noteworthy observation here - in the *MovieLens 100k* dataset, those users or items that contribute positively to fairness end up negatively impacting accuracy. This brings to light a possible balancing act between fairness and accuracy that needs to be managed.

Method	Random-User	Random-Item	Active-User	Inactive-User	Active-Item	Inactive-Item
Toys and Games						
ACFR - $\text{Item}_i(-)(\Psi_{user})$	0.0106	0.3239	0.001	0.1198	0.0025	0.0862
ACFR - $\text{Item}_i(+)(\Psi_{item})$	0.001	0.001	0.001	0.001	0.001	0.001
Beauty						
ACFR - $\text{Item}_i(-)(\Psi_{user})$	0.001	0.035	0.001	0.001	0.0207	0.9
ACFR - $\text{User}_i(+)(\Psi_{item})$	0.002	0.001	0.0713	0.001	0.001	0.001
MovieLens 100k						
ACFR - $\text{User}_u(+)(\Psi_{user})$	0.001	0.0134	0.001	0.001	0.001	0.0021
ACFR - $\text{User}_i(+)(\Psi_{item})$	0.001	0.001	0.006	0.001	0.001	0.001

Table 6. The p-value results of Tukey's HSD ($\alpha' = 0.05$).

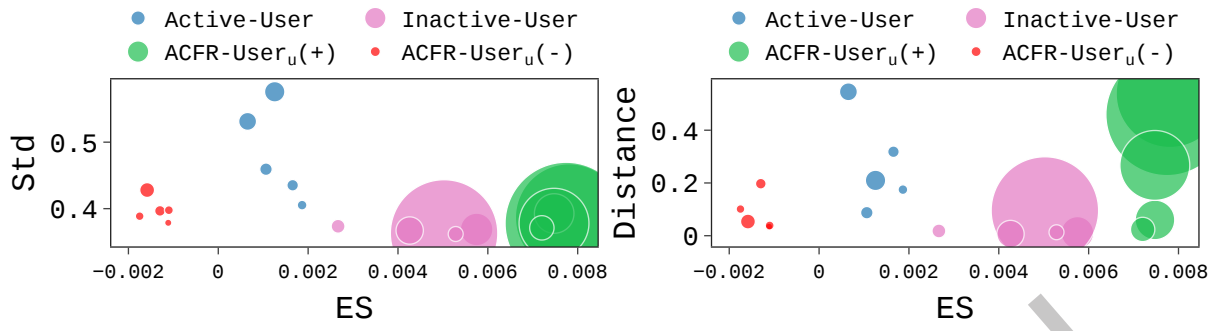
On the other hand, in the *Toys and Games* and *Beauty* datasets, we observe a harmonious relation where users or items positively affecting fairness also enhance accuracy. This underscores the fact that the trade-off between fairness and accuracy, while prevalent, is not a steadfast rule and can be dataset-specific. This finding paves the way for a more intricate understanding and strategic handling of the balance between fairness and accuracy in recommendation systems.

5.5.7 Determining the Significance of Differences in Fairness: Tukey's HSD. The relative effectiveness of the proposed *ACFR* and the baselines in terms of fairness is examined through the application of Tukey's Honest Significant Differences (HSD) [1]. The corresponding p-values are laid out in table 6. *ACFR* is deemed to outperform the baselines significantly if the resulting p-value falls below 0.05 ($\alpha' = 0.05$) with respect to recommendation fairness.

5.6 Analysis on Interactions of Selected Users

To shed light on why selected users and items play a crucial role in recommendation fairness, we analyze the relationship between the distribution of interactions of selected users/items and their ES. Due to the page limit, we only show the investigation results of explanations in terms of user fairness by methods selecting users. We first plot the standard deviation of the selected users' interactions against their ES generated in terms of user fairness on the Beauty dataset (fig. 1a). Different colors represent different methods for selecting users to explain fairness among individual users. The results in fig. 1a show that while the standard deviation of interactions is similar for red dots (*ACFR* - $\text{User}_u(-)$) and green dots (*ACFR* - $\text{User}_u(+)$), the ES of green dots is significantly larger than that of red dots. Blue dots (Active-users) have a unique distribution, while pink dots (Inactive-users) have a distribution similar to red dots and green dots. Since the red dots, pink dots and green dots show similar standard deviation in fig. 1a, we further examine the relationship between the ES of selected users and the euclidean distance between their interactions and the average user (who is denoted by the average of all items' interactions from all users). The results in fig. 1b show that the interactions of users selected by *ACFR* - $\text{User}_u(+)$ (green dots) are different from most users, while red dots and pink dots have a relatively smaller distance from the average. This highlights the significant difference between users selected by *ACFR* - $\text{User}_u(+)$ and *ACFR* - $\text{User}_u(-)$ despite their similar interaction distributions.

To understand why pink dots have a higher ES than red dots despite having a similar standard deviation (fig. 1a) and distance from the average (fig. 1b), we calculate the mean squared error between $\tilde{\mathbf{r}}$ and $\hat{\tilde{\mathbf{r}}}$ for each selected user (as defined in eq. (7)). This is to determine which users can provide more information to the pre-trained



(a) Standard Deviation VS. ES

(b) Euclidean Distance VS. ES

Fig. 1. Analysis on interactions of selected users on the Beauty dataset.

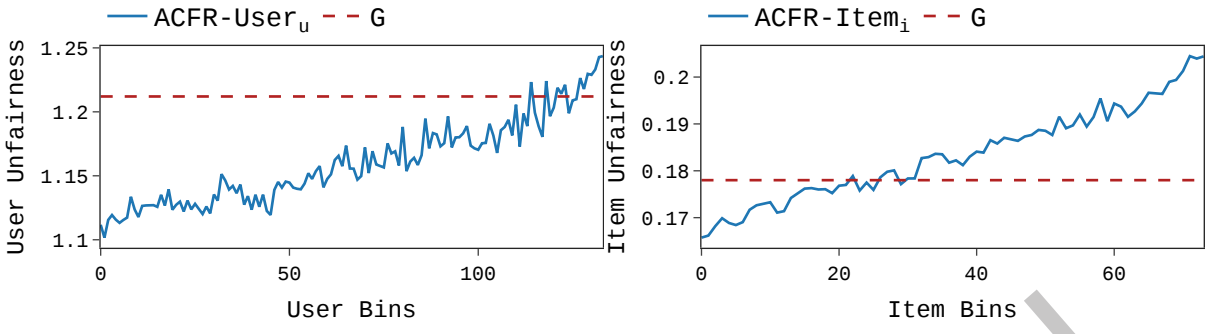
Method	Rand-U	Rand-I	Act-U	Inact-U	Act-I	Inact-I	*ACFR* - User$_u$(+)
User Unfairness (N=10)							
DeepFM	1.191	1.166	1.254	1.179	1.200	1.187	**1.149**
FwFM	1.200	1.166	1.245	1.180	1.201	1.190	**1.148**
FNN	1.175	1.195	1.250	1.181	1.209	1.200	**1.150**

Table 7. Effect of different imputation in Toys and Games.

model G. The larger the error, the more different information is introduced by the interactions of the selected user, and the more likely the user will impact fairness (improve fairness by eq. (8)). We display the error as the size of dots in fig. 1a and fig. 1b, with larger dots indicating larger errors. In both figures, pink dots and green dots have relatively larger errors, which suggests that larger errors have a larger impact on fairness. This can also explain why pink dots have a higher ES than red dots even they have the similar standard deviation and distance from the average. Note that the interactions analyzed in this investigation comprise both historical and imputed ones, including the calculation of the average of all items' interactions.

5.7 Analysis on Different Imputation

In order to thoroughly assess the performance of the proposed system when using different imputation models, we conducted an evaluation that included two additional recommendation models, FwFM [41] and FNN [75], which were represented as $\mathcal{G}$. By comparing these models, we aimed to gain a better understanding of how the proposed *ACFR* operates in relation to other approaches within the field. The user unfairness results for both the baselines and the proposed *ACFR* are presented in table 7. Upon analyzing the data provided in table 7, it becomes evident that the proposed *ACFR* consistently outperforms the baselines, regardless of the specific imputation models employed. This observation is highly significant, as it suggests that the proposed *ACFR* is not sensitive to the imputed interactions. Consequently, we can infer that the *ACFR* is highly adaptable and versatile.



(a) *ACFR* - User$_u$ (b) *ACFR* - Item$_i$

Fig. 2. The positive and negative effects of users and items on fairness in the Beauty dataset.

5.8 Positive and Negative Impact on Fairness

We evaluated the influence of certain users and items on fairness in the recommendation model **G** using *ACFR* - User$_u$ and *ACFR* - Item$_i$ on the Beauty dataset. We calculated explanation scores for each user and item, ranked them in descending order, and grouped them into bins of 10. The bin on the far left of the x-axis represents users or items with the highest ES, while the bin on the far right represents users or items with the lowest ES. The blue line represents the performance of user or item unfairness after retraining with imputed interactions for users or items in each bin. The red dashed line shows the performance of the original recommendation model **G** in terms of user or item unfairness. The results showed that users starting from the bin immediately to the right of the 115th bin negatively impacted user fairness (as shown in fig. 2a), while users in the bins to the left of the 115th bin had a positive impact. Similarly, items in bins to the left of the 30th bin positively impacted item fairness (as shown in fig. 2b), while items in bins to the right of the 30th bin had a negative impact. In conclusion, the proposed *ACFR* is capable of identifying not just users and items that improve fairness, but also those that negatively impact it.

5.9 More Recent Dataset

In this section, we assess the *ACFR* framework's performance using the MovieLens 25M dataset to demonstrate its applicability to more contemporary datasets. Due to the large size of users (162,000) and items (62,000), without loss of generality, we randomly selected a subset of 2,000 users and 2,000 items for our experiment. Furthermore, to examine their influence on user fairness (Initial User Unfairness of **G**: 1.606), we run three separate checks by randomly selecting 150 items three times. The outcomes of these analyses are depicted in fig. 3. The x-axis in fig. 3 categorizes the items into bins, with each bin containing 10 items. These items are sorted according to their Explanation Scores (*ES*), with '1' on the x-axis indicating the top 10 items, '6' indicating the top 60, and so forth. The leftmost graph in fig. 3 demonstrates that nearly all selected items positively impact user fairness, as indicated by the consistent decrease in user unfairness with the addition of more items. Conversely, the middle and right graphs reveal that the selected items exhibit both positive and negative effects on user fairness.

5.10 Decision Tree Structure

In this section, we employ decision trees to illuminate the influence different categories of users and items exert on the fairness of our recommendations, as shown in fig. 4 and fig. 5. We draw on data from the *MovieLens 100k* dataset to construct these decision trees.

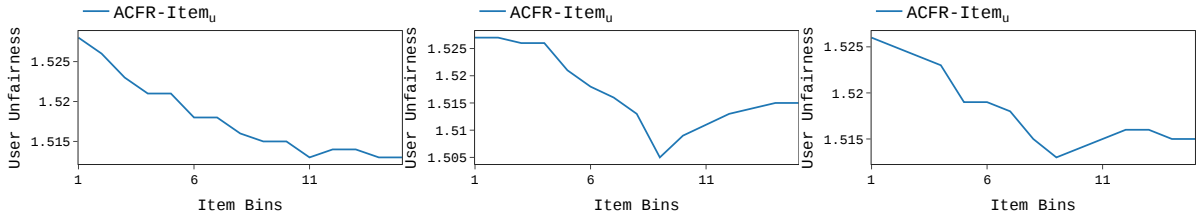


Fig. 3. The effects of selected items on user fairness in the MovieLens 25M dataset.

Methods	RMSE			User Unfairness		
	N=10	N=20	N=30	N=10	N=20	N=30
The Model G		0.959			1.280	
ACFR - User$_u(-)^{\Delta\tilde{\omega}-}$	0.955	0.953	0.952	1.263	1.265	1.257
ACFR - User$_u(-)$	0.956	0.955	0.954	1.269	1.275	1.272
ACFR - User$_u(+)^{\Delta\tilde{\omega}-}$	0.939	0.939	0.939	1.152	1.150	1.149
ACFR - User$_u(+)$	**0.939**	**0.939**	**0.939**	**1.149**	**1.148**	**1.147**

Table 8. Results of ablation analysis on the performance of RMSE and user fairness on Toys and Games dataset. The best results are bold-faced.

On the user side, we discovered, for instance, that users with the occupation of 'Scientist' have the highest average *ES* of 96.97, indicating a significant impact on user fairness. At the other end, 'Female Students' demonstrate the lowest average *ES* of 19.49, implying a lesser impact. Similarly, we uncovered that certain genres of movies tend to influence item fairness either positively or negatively.

Practically speaking, these insights can be instrumental in multiple ways. For instance, recommender systems could prioritize users or items identified as potentially causing fairness discrepancies for fairness improvement interventions. Alternatively, these decision trees can guide the design of new recommendation algorithms that inherently mitigate unfairness based on user or item categories. In summary, the understanding gleaned from this analysis can be an invaluable resource in shaping strategies to enhance recommendation fairness.

5.11 Ablation Analysis

We conducted an ablation analysis to determine if it is necessary to minimize the changes in $\tilde{\omega}$ of **G** while optimizing the counterfactual optimization problem (as defined in eq. (8)). We tested a different setting for *ACFR* - User$_u$, referred to as *ACFR* - User$_u^{\Delta\tilde{\omega}-}$, which represents removing the second term $||\Delta\tilde{\omega}||_2^2$ in eq. (8). The results in table 8 show that *ACFR* - User$_u(+)$ performs better than *ACFR* - User$_u(+)^{\Delta\tilde{\omega}-}$ in terms of user fairness, while *ACFR* - User$_u(-)$ has higher user unfairness than *ACFR* - User$_u(-)^{\Delta\tilde{\omega}-}$, indicating that the term $||\Delta\tilde{\omega}||_2^2$ is important in determining the amount of effort a user needs to make to improve fairness.

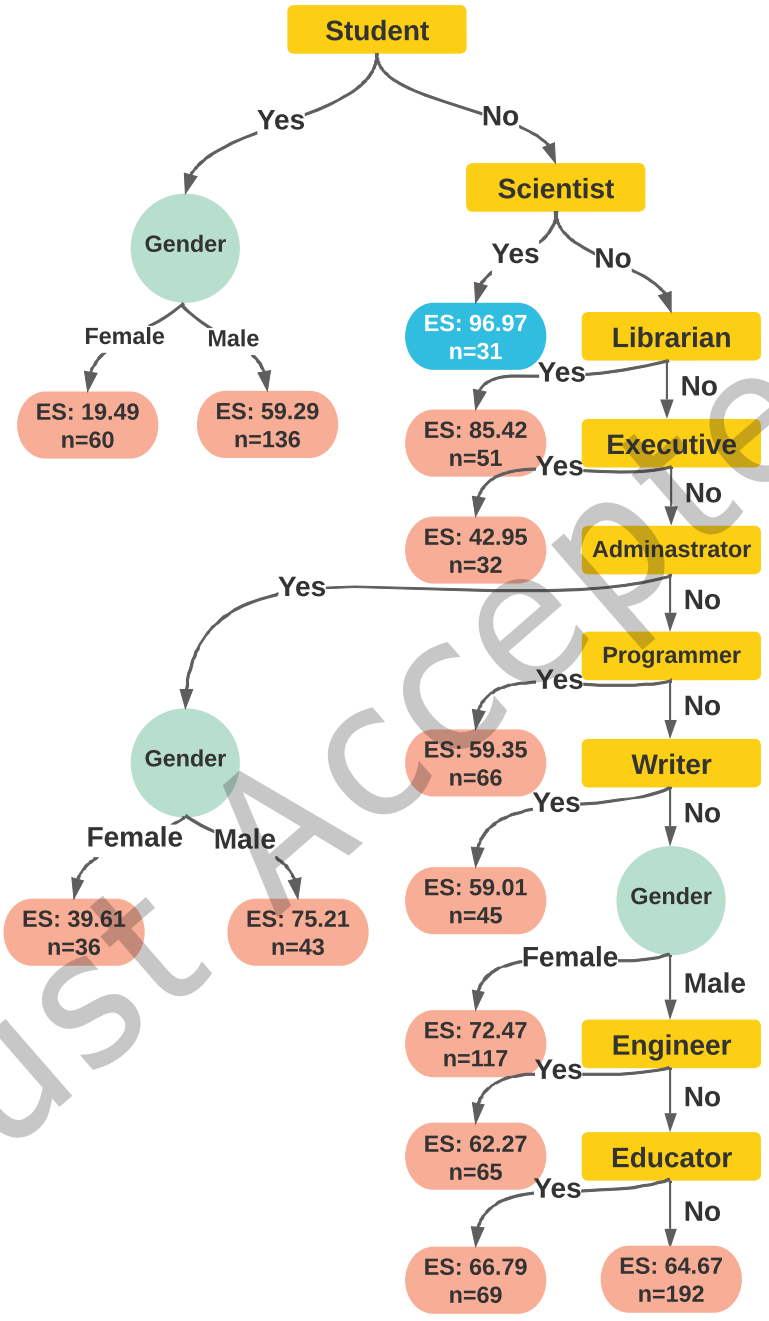


Fig. 4. The relationship between user profiles and ES produced by *ACFR* - User$_u$ on the MovieLens dataset.

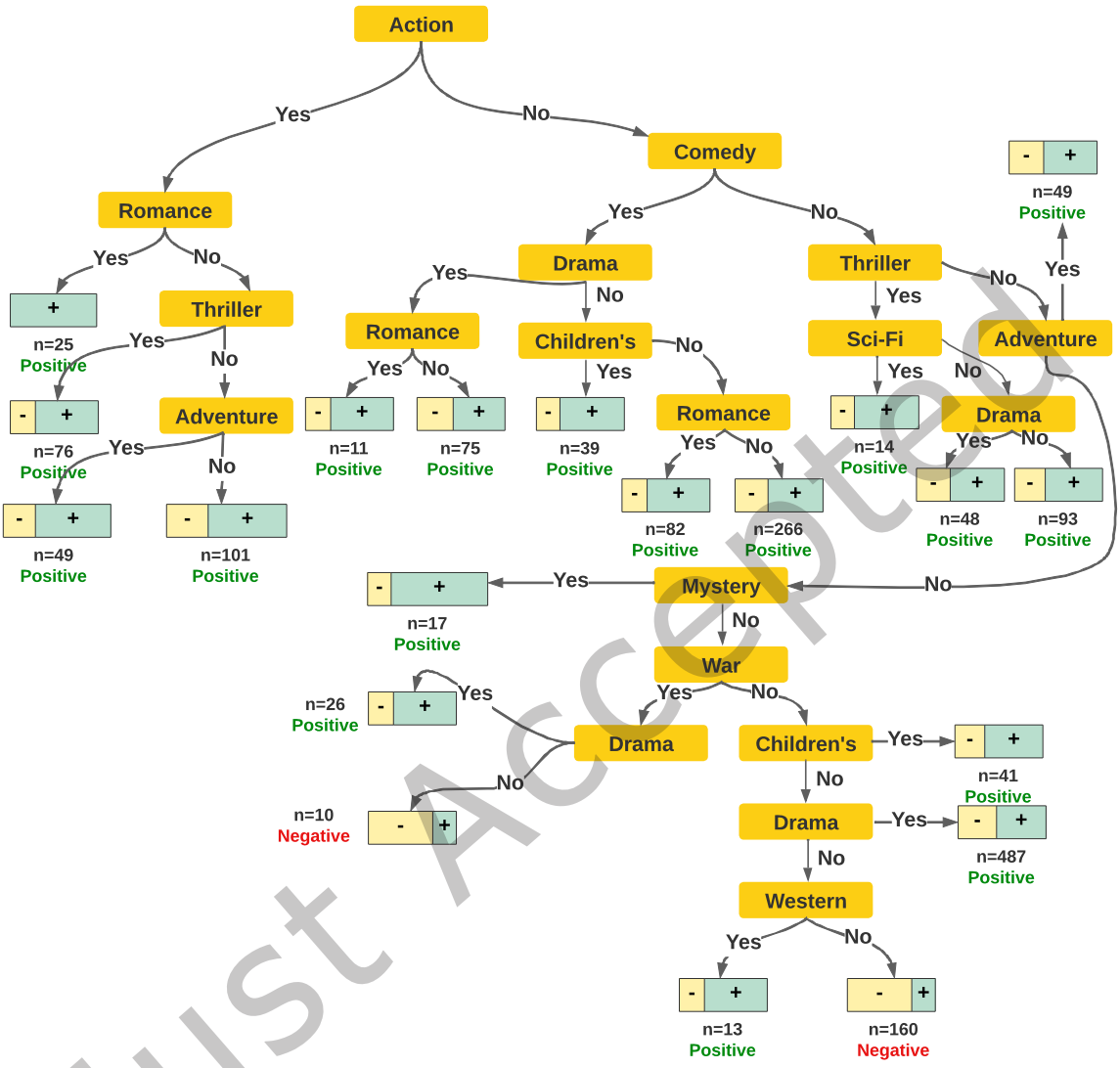
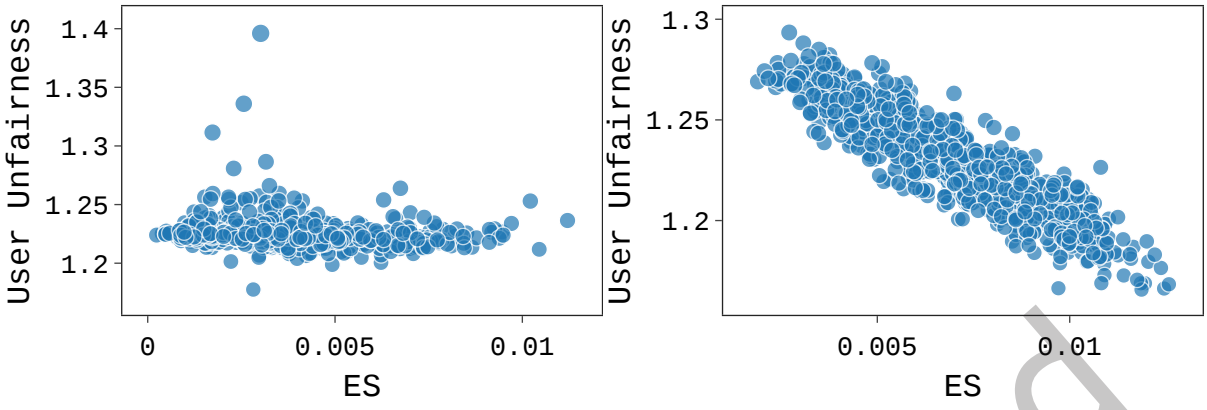


Fig. 5. The relationship between item profiles and fairness impact on the MovieLens dataset.

5.12 Erasing-based Counterfactual Analysis

In our research, we also employed an erasing-based counterfactual analysis strategy. This approach involved the removal of current interactions of specific users/items rather than the addition of unknown interactions, the latter being a characteristic of the adding-based method utilized in *ACFR*. The erasing-based method is comprised of the following procedures:

(1) training the recommendation model **G** following eq. (4).



(a) Erasing-based method.

(b) Adding-based method.

Fig. 6. The relationship between user unfairness and ES generated on the Toys and Games dataset.

(2) for each user, minimizing the objective function as defined in eq. (8)). Unlike the adding-based method discussed in section 4.2 where all weights of model **G** are affected, this method only updates the weights related to the current user, those users who have interactions with the same items of the current user, and items interacted with by the current user.
(3) calculating the ES for the erased user following section 4.3.

The outcomes, as depicted in fig. 6a and fig. 6b, demonstrate the relationship between user unfairness and the corresponding ES of said user. For clarity, in fig. 6a and fig. 6b, each dot represents a single, unique user. The x-axis corresponds to the ES of the user, and the y-axis denotes user unfairness, thereby indicating the shift in fairness after either eliminating the specific user's interactions or adding imputed interactions for this user. This graphic illustration allows us to understand how modifications to each user's interactions can influence the total fairness of the recommender system.

The expectation in fig. 6a was that the value of user unfairness would be lower when removing users with higher ES, but no clear trend can be observed. In contrast, the trend is clear in fig. 6b, which indicates the relationship between the user unfairness that is based on predictions produced by pre-trained **G** after imputing missing interactions of each user and the corresponding ES of the user generated by *ACFR* - User. This suggests that the erasing-based method does not produce meaningful explanations, while the adding-based method does.

6 DISCUSSION

This research was undertaken with the goal of advancing our understanding of fairness in recommender systems. With a keen focus on individual user-item interactions, we delved deep into exploring how fairness manifests at this granular level and what it implies for the overall integrity of a recommender system.

In the realm of fairness and explainability in recommendations, our work threads a nuanced path, filling gaps that have hitherto been overlooked. While the existing literature offers methodologies to ensure fairness in recommendations and explainability for such recommendations, few have embarked on the journey of explaining the fairness of these recommendations themselves. This dichotomy is what our research bridges. Several seminal works, like those of [9] and [18], ventured into explaining the root causes of unfairness. Yet, these explanations were often couched in terms of features and lacked the user-item interaction perspective. Our emphasis on

individual interactions as the foundation of recommender systems is thus a significant departure from the established paradigms.

Upon examining the recent perspectives paper by Sun [58], we observed intriguing parallels between our vision for improved explanation systems and his call for refinement in the evaluation domain. Sun stresses the need for more nuanced evaluation methodologies in information retrieval, emphasizing the importance of context, user behavior, and long-term impact. These aspects resonate with our own perspective of diving deep into the granularity of individual user-item interactions and examining fairness from this lens. It's imperative to note that while Sun's work centers on a broader call for comprehensive evaluation frameworks, our research delves into the specifics of fairness explanations within recommender systems. However, the shared ethos of a deeper, more contextual understanding underpins both. Our work can be seen as a practical realization of some of the tenets Sun proposes, albeit in a more focused domain. As recommender systems evolve and adapt, it's plausible that the fairness explanations we propound become integral to the advanced evaluation methodologies that Sun envisions.

7 CONCLUSION

In the intricate world of recommender systems, fairness has become a crucial focal point. This study embarked on an ambitious journey to navigate this complex terrain, addressing four pivotal research questions.

Our first inquiry revolved around understanding the role of individual user-item interactions in shaping, and at times distorting, the fairness of recommender systems. Through our research, we found that these granular interactions can significantly influence the overall perception of fairness, often amplifying biases inherent in the system, a finding that emphasizes the importance of treating each user and item as unique entities.

In answering the second research question, we delved into creating a methodological framework that prioritizes these individual user-item interactions. Our objective was to unearth the underlying sources of unfairness. By meticulously dissecting these interactions, we provided deeper insights into where biases may originate and how they proliferate within the system.

For our third research question, we introduced the Adding-based Counterfactual Fairness Reasoning (*ACFR*) and elucidated its advantages over traditional fairness methodologies. Notably, *ACFR* transcends conventional approaches by imputing unknown interactions, adeptly navigating around the 'no gradients' issue often encountered during training. This nuanced approach not only bolsters the robustness of the model but significantly enhances its clarity and interpretability.

Lastly, our fourth inquiry aimed at assessing the empirical strength of the *ACFR* approach. Our findings demonstrate that *ACFR* offers a more compelling narrative in explaining recommendation fairness, outperforming existing methods in both precision and depth of explanation.

To conclude, this work shines an analytical spotlight on recommendation fairness, offering a fresh perspective through the lens of *ACFR*. By addressing the unique challenges presented by user-item interactions and proffering an innovative methodology, we aim to pave the way for more equitable and transparent recommender systems in the future.

REFERENCES

[1] Hervé Abdi and Lynne J Williams. 2010. Tukey's honestly significant difference (HSD) test. *Encyclopedia of research design* 3, 1 (2010), 1–5.

[2] Behnoush Abdollahi and Olfa Nasraoui. 2018. Transparency in fair machine learning: the case of explainable recommender systems. In *Human and machine learning*. Springer, 21–35.

[3] Himan Abdollahpouri and Robin Burke. 2019. Multi-stakeholder recommendation and its connection to multi-sided fairness. *arXiv preprint arXiv:1907.13158* (2019).

[4] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. 2005. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information systems (TOIS)* 23, 1 (2005), 103–145.

[5] Gediminas Adomavicius and Alexander Tuzhilin. 2005. 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 (2005), 734–749.

[6] Darius Afchar, Alessandro B Melchiorre, Markus Schedl, Romain Hennequin, Elena V Epure, and Manuel Moussallam. 2022. Explainability in Music Recommender Systems. *arXiv preprint arXiv:2201.10528* (2022).

[7] Abdul Basit Ahanger, Syed Wajid Aalam, Muzafar Rasool Bhat, and Assif Assad. 2022. Popularity bias in recommender systems-a review. In *International Conference on Emerging Technologies in Computer Engineering*. Springer, 431–444.

[8] Ting Bai, Ji-Rong Wen, Jun Zhang, and Wayne Xin Zhao. 2017. A neural collaborative filtering model with interaction-based neighborhood. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1979–1982.

[9] Tom Begley, Tobias Schwedes, Christopher Frye, and Ilya Feige. 2020. Explainability for fair machine learning. *arXiv preprint arXiv:2010.07389* (2020).

[10] Vito Bellini, Angelo Schiavone, Tommaso Di Noia, Azzurra Ragone, and Eugenio Di Sciascio. 2018. Knowledge-aware autoencoders for explainable recommender systems. In *Proceedings of the 3rd workshop on deep learning for recommender systems*. 24–31.

[11] Sarah Bird, Krishnaram Kenthapadi, Emre Kiciman, and Margaret Mitchell. 2019. Fairness-aware machine learning: Practical challenges and lessons learned. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 834–835.

[12] Jesús Bobadilla, Raúl Lara-Cabrera, Ángel González-Prieto, and Fernando Ortega. 2020. Deepfair: deep learning for improving fairness in recommender systems. *arXiv preprint arXiv:2006.05255* (2020).

[13] Xu Chen, Hanxiong Chen, Hongteng Xu, Yongfeng Zhang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2019. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 765–774.

[14] Yashar Deldjoo, Alejandro Bellogin, and Tommaso Di Noia. 2021. Explaining recommender systems fairness and accuracy through the lens of data characteristics. *Information Processing & Management* 58, 5 (2021), 102662.

[15] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 275–284.

[16] Consistent Distinguishability. 2013. A Theoretical Analysis of Normalized Discounted Cumulative Gain (NDCG) Ranking Measures. (2013).

[17] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, et al. 2020. Fairness-aware explainable recommendation over knowledge graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 69–78.

[18] Yingqiang Ge, Juntao Tan, Yan Zhu, Yinglong Xia, Jiebo Luo, Shuchang Liu, Zuohui Fu, Shijie Geng, Zelong Li, and Yongfeng Zhang. 2022. Explainable fairness in recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 681–691.

[19] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*. 2221–2231.

[20] Mustansar Ali Ghazanfar and Adam Prugel. 2013. The advantage of careful imputation sources in sparse data-environment of recommender systems: Generating improved svd-based recommendations. *Informatica* 37, 1 (2013).

[21] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy, and Gerhard Weikum. 2020. PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 196–204.

[22] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).

[23] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[24] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[25] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*. 507–517.

[26] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 5–53.

[27] Won-Seok Hwang, Shaoyu Li, Sang-Wook Kim, and Kichun Lee. 2018. Data imputation using a trust network for recommendation via matrix factorization. *Computer Science and Information Systems* 15, 2 (2018), 347–368.

[28] Rob J Hyndman and Anne B Koehler. 2006. Another look at measures of forecast accuracy. *International journal of forecasting* 22, 4 (2006), 679–688.

[29] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[30] Jiangli Jiao, Xueying Zhang, Fenglian Li, and Yan Wang. 2019. A novel learning rate function and its application on the SVD++ recommendation algorithm. *IEEE Access* 8 (2019), 14112–14122.
[31] Nima Kordzadeh and Maryam Ghasemaghaei. 2022. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems* 31, 3 (2022), 388–409.
[32] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
[33] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User fairness in recommender systems. In *Companion Proceedings of the The Web Conference 2018*. 101–102.
[34] Jie Li, Yongli Ren, and Ke Deng. 2022. FairGAN: GANs-Based Fairness-Aware Learning for Recommendations with Implicit Feedback. In *Proceedings of the ACM Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 297–307. https://doi.org/10.1145/3485447.3511958
[35] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented fairness in recommendation. In *Proceedings of the web conference 2021*. 624–632.
[36] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.
[37] Hai Liu, Chao Zheng, Duantengchuan Li, Xiaoxuan Shen, Ke Lin, Jiazhang Wang, Zhen Zhang, Zhaoli Zhang, and Neal N Xiong. 2021. EDMF: Efficient deep matrix factorization with review feature learning for industrial recommender system. *IEEE Transactions on Industrial Informatics* 18, 7 (2021), 4361–4371.
[38] Ninghao Liu, Yong Ge, Li Li, Xia Hu, Rui Chen, and Soo-Hyun Choi. 2020. Explainable recommender systems via resolving learning representations. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 895–904.
[39] Christoph Molnar. 2020. *Interpretable machine learning*. Lulu. com.
[40] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 429–438.
[41] Junwei Pan, Jian Xu, Alfonso Lobos Ruiz, Wenliang Zhao, Shengjun Pan, Yu Sun, and Quan Lu. 2018. Field-weighted factorization machines for click-through rate prediction in display advertising. In *Proceedings of the 2018 World Wide Web Conference*. 1349–1357.
[42] Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, and Fei Wang. 2021. Explaining algorithmic fairness through fairness-aware causal path decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1287–1297.
[43] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. 2020. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of the web conference 2020*. 1194–1204.
[44] Georgina Peake and Jun Wang. 2018. Explanation mining: Post hoc interpretability of latent factor models for recommendation systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2060–2069.
[45] David MW Powers. 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061* (2020).
[46] Amifa Raj and Michael D Ekstrand. 2022. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 726–736.
[47] 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*. 231–239.
[48] Yongli Ren, Gang Li, Jun Zhang, and Wanlei Zhou. 2012. The efficient imputation method for neighborhood-based collaborative filtering. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. 684–693.
[49] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).
[50] Lance J Rips and Brian J Edwards. 2013. Inference and explanation in counterfactual reasoning. *Cognitive Science* 37, 6 (2013), 1107–1135.
[51] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion proceedings of the 2019 world wide web conference*. 553–562.
[52] Joseph L Schafer and John W Graham. 2002. Missing data: our view of the state of the art. *Psychological methods* 7, 2 (2002), 147.
[53] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *Proceedings of the eleventh ACM conference on recommender systems*. 297–305.
[54] Lloyd S Shapley. 1997. A value for n-person games. *Classics in game theory* 69 (1997).
[55] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.
[56] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. *Advances in neural information processing systems* 32 (2019).
[57] Nasim Sonboli, Jessie J Smith, Florencia Cabral Berenfus, Robin Burke, and Casey Fiesler. 2021. Fairness and transparency in recommendation: The users' perspective. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 274–279.

[58] Aixin Sun. 2023. Take a Fresh Look at Recommender Systems from an Evaluation Standpoint. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2629–2638.
[59] Juntao Tan, Shijie Geng, Zuohui Fu, Yingqiang Ge, Shuyuan Xu, Yunqi Li, and Yongfeng Zhang. 2022. Learning and evaluating graph neural network explanations based on counterfactual and factual reasoning. In *Proceedings of the ACM Web Conference 2022*. 1018–1027.
[60] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. 2021. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1784–1793.
[61] Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. 2021. Counterfactual Explanations for Neural Recommenders. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1627–1631.
[62] Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2019. Multi-task feature learning for knowledge graph enhanced recommendation. In *The world wide web conference*. 2000–2010.
[63] Ningxia Wang and Li Chen. 2021. User bias in beyond-accuracy measurement of recommendation algorithms. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 133–142.
[64] Shijie Wang, Guiling Sun, and Yangyang Li. 2020. SVD++ recommendation algorithm based on backtracking. *Information* 11, 7 (2020), 369.
[65] Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. 2018. A reinforcement learning framework for explainable recommendation. In *2018 IEEE international conference on data mining (ICDM)*. IEEE, 587–596.
[66] Xiang Wang, Tinglin Huang, Dingxian Wang, Yancheng Yuan, Zhenguang Liu, Xiangnan He, and Tat-Seng Chua. 2021. Learning intents behind interactions with knowledge graph for recommendation. In *Proceedings of the Web Conference 2021*. 878–887.
[67] Xiang Wang, Yaokun Xu, Xiangnan He, Yixin Cao, Meng Wang, and Tat-Seng Chua. 2020. Reinforced negative sampling over knowledge graph for recommendation. In *Proceedings of the web conference 2020*. 99–109.
[68] Yifan Wang, Weizhi Ma, Min Zhang*, Yiqun Liu, and Shaoping Ma. 2022. A survey on the fairness of recommender systems. *ACM Journal of the ACM (JACM)* (2022).
[69] Weiwei Xia, Liang He, Junzhong Gu, and Keqin He. 2009. Effective collaborative filtering approaches based on missing data imputation. In *2009 Fifth International Joint Conference on INC, IMS and IDC*. IEEE, 534–537.
[70] Shuyuan Xu, Yunqi Li, Shuchang Liu, Zuohui Fu, Yingqiang Ge, Xu Chen, and Yongfeng Zhang. 2021. Learning causal explanations for recommendation. In *The 1st International Workshop on Causality in Search and Recommendation*.
[71] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 1–6.
[72] Baolin Yi, Xiaoxuan Shen, Hai Liu, Zhaoli Zhang, Wei Zhang, Sannyuya Liu, and Naixue Xiong. 2019. Deep matrix factorization with implicit feedback embedding for recommendation system. *IEEE Transactions on Industrial Informatics* 15, 8 (2019), 4591–4601.
[73] Xiaofeng Yuan, Lixin Han, Subin Qian, Guoxia Xu, and Hong Yan. 2019. Singular value decomposition based recommendation using imputed data. *Knowledge-Based Systems* 163 (2019), 485–494.
[74] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000* (2021).
[75] Weinan Zhang, Tianming Du, and Jun Wang. 2016. Deep Learning over Multi-field Categorical Data: –A Case Study on User Response Prediction. In *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20–23, 2016. Proceedings 38*. Springer, 45–57.
[76] Yongfeng Zhang, Xu Chen, et al. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101.
[77] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 83–92.