Fairness in Recommendation: A Survey

YUNQI LI, Rutgers University, USA, yunqi.li@rutgers.edu
HANXIONG CHEN, Rutgers University, USA, hanxiong.chen@rutgers.edu
SHUYUAN XU, Rutgers University, USA, shuyuan.xu@rutgers.edu
YINGQIANG GE, Rutgers University, USA, yingqiang.ge@rutgers.edu
JUNTAO TAN, Rutgers University, USA, juntao.tan@rutgers.edu
SHUCHANG LIU, Rutgers University, USA, shuchang.liu@rutgers.edu
YONGFENG ZHANG, Rutgers University, USA, yongfeng.zhang@rutgers.edu

Recently, there has been growing attention on fairness considerations in recommender systems with more and more literature on approaches to promote fairness in recommendation. However, the studies are rather fragmented and lack a systematic organization, thus making it difficult to penetrate for new researchers to the domain. This motivates us to provide a systematic survey of existing works on fairness in recommendation. This survey focuses on the foundations for fairness in recommendation literature. It first presents a brief introduction about fairness in basic machine learning tasks such as classification and ranking in order to provide a general overview of fairness research, as well as introduce the more complex situations and challenges that need to be considered when studying fairness in recommender systems. After that, the survey will introduce fairness in recommendation with a focus on the taxonomies of current fairness definitions, the typical techniques for improving fairness, as well as the datasets for fairness studies in recommendation. The survey also talks about the challenges and opportunities in fairness research with the hope of promoting the fair recommendation research area and beyond.

CCS Concepts: • **Computing methodologies → Machine learning**; • **Information systems → Recommender systems**.

Additional Key Words and Phrases: Fairness; Recommender System; Machine Learning

1 INTRODUCTION

Recommender systems are playing an essential role in the era of information explosion. They help users to gain access to the information they are interested in and benefit sellers or producers to increase their exposure so as to gain profits. Since recommender systems are associated with multi-sided benefits by deciding what information to be delivered, fairness becomes an especially critical problem that requires attention. However, in the literature on fairness research, researchers have shown that the recommender system, as a highly data-driven application, can suffer from various unfairness issues and may harm the benefits of multiple stakeholders [1, 15, 29, 73, 107, 112, 114, 117, 202]. For example, an unfair job recommender system may exhibit racial or gender discrimination by disproportionately recommending low-payment jobs to certain (protected) user groups; e-commerce recommender systems may disproportionately recommend the products from popular sellers while providing very limited exposure opportunity to other equally qualified but less popular items; news recommender systems may over recommend specific political ideologies over others (sometimes due to echo chambers) which may manipulate user's opinions. To improve the satisfaction of various participants, it is important to solve the unfairness issue in recommender systems to build a positive and sustainable ecosystem. In recent years, works on fair recommendation have come to the fore and received increasing attention. The goal of this survey is to synthesize the current state of fairness in recommendation and to inspire more future work in this area.

We begin the survey with a brief introduction of fairness in machine learning to provide the readers a general and basic background knowledge of fairness research. We talk about the main causes of unfairness, the typical methods for achieving fairness, and the diverse fairness considerations in machine learning. Though focusing on fairness in recommendation, we first talk about certain fairness concerns and representative methods in classification and ranking tasks due to the following considerations: first, we would like to take fairness in basic machine learning tasks such as classification and ranking as examples to help readers better understand the key concepts introduced in the previous research; second, before studying fairness in recommendation, the first endeavor to achieve fairness in the community is to develop fair classifiers and rankers. Therefore, these two areas provide insightful knowledge and can be enlightening for the researches in fair recommendation since recommendation problems can usually be formulated as classification or ranking tasks.

However, promoting fairness in recommendation can face unique challenges. For example, recommender systems often consist of multiple models to balance multiple goals; recommender systems are dynamic and need to consider long-term benefits; the extreme data sparsity can also bring additional difficulties for model learning and evaluating. What's more, unlike fairness in classification and ranking problems that usually consider a single-side fairness requirement, the concept of fairness in the recommendation has been extended to multiple-stakeholders [29] and consequently make the fairness problem more challenging. In the main body of the survey, we introduce fairness in recommendation from four perspectives: *taxonomy*, *techniques*, *datasets* and *open challenges*. As aforementioned, recommender systems consider complicated application scenarios so that researchers view unfairness issues from very different perspectives. This motivates us to provide a systematical taxonomy of fairness notions in recommendation to help readers have a clear understanding of how to consider unfairness problems in recommendation scenario. After that, we introduce the typical techniques for promoting fairness in recommendation to help explain how to alleviate the various unfairness concerns. Additionally, we collect and present several publicly available datasets with user sensitive features to benefit future fairness studies in recommendation. Last but not the least, we summarize open challenges and opportunities in fair recommendation research to suggest future directions of this area.

In the past few years, a number of surveys talking about fairness and bias in general machine learning have been published [34, 35, 129, 141, 175], but they usually focus on or give examples of the fairness works in binary classification tasks. Some other surveys pay attention to one certain field in machine learning or fairness works such as [165], which provides a comprehensive survey to help readers gain insights into fairness-aware federated learning; [213] focuses on the existing literature on the fairness of data-driven sequential decision-making area; and [125] reviews an exhaustive list of causal-based fairness notions and study their applicability in real-world scenarios. A few surveys provide an overview of fairness in ranking tasks [144, 209]. Recommendation algorithms can usually be considered as a type of ranking algorithm. However, existing works on fair ranking usually only consider unfairness issue from the item perspective, such as the item exposure fairness, while the concept of fairness in recommender systems can be more complicated and needs to be thought about from multiple sides [29]. Pitoura et al. [148] talk about fairness in both ranking and recommendation. Though covering a brief introduction about fairness in classification and ranking, our survey pays specific attention to organizing the foundations of fairness in recommendation including a more comprehensive taxonomy of fairness notions proposed in recommendation problem, task-specific techniques for promoting recommendation fairness, as well as datasets specially for research in recommendation. Moreover, Chen et al.[43] provide a survey on bias and debias in recommender system, which covers a part of content about fairness in recommendation. However, most of the works on bias in recommendation focus on improving the recommendation accuracy or robustness in out-of-distribution scenario through debiasing methods instead of promoting fairness. Our survey differs from previous work as we concentrate on the unfairness issue in recommendation and provide a broader and more in-depth introduction.

The remaining of this survey is organized as follows. We provide a general introduction about fairness in machine learning in Section 2. Section 3 and Section 4 briefly talk about fairness studies in classification and ranking tasks, respectively. In Section 5, we introduce fairness in recommendation. We discuss some important challenges and opportunities of fairness research in recommendation in Section 6, and Section 7 concludes the survey.

2 FAIRNESS IN MACHINE LEARNING

In this section, we provide basic background knowledge of fairness in machine learning, including the causes of unfairness, fair learning methods, diverse fairness definitions, as well as the corresponding evaluation metrics.

2.1 The Causes of Unfairness

Unfairness in machine learning is usually caused by various forms of biases. A list of different sources of biases with their corresponding definitions in machine learning has been provided in several previous surveys [34, 129, 140, 170]. It is out of the scope of this survey to list and discuss all of the types of biases in machine learning, nevertheless, we want to talk about the main sources of biases from a higher level. Generally speaking, as the two fundamental components of machine learning system, training data and learning algorithm are also the main sources of biases which can lead to unfair results in machine learning tasks. Therefore, among all kinds of biases, we categorize them into two types: *Data Bias* and *Algorithmic Bias*.

2.1.1 ***Data Bias****.* Most of the biases in machine learning lie in the data itself. The bias in data can come from the processes of data generation, data collection, and data storage. Here we roughly categorize the data bias into *Statistical Bias* and *Pre-existing Bias*.

Statistical Bias [34]. The statistical bias usually arises from the process of data collection and storage. It occurs when there are flaws in the experimental design or data collection process and thus the data is not the true representation of the population. For example, the data is not randomly selected from the full population, is wrongly recorded, or is systematically missing, resulting in the lack of population diversity or other anomalies. Statistical bias can easily lead to systematic differences between the true parameters and the estimated statistics of a population.

Pre-existing Bias [170]. Even if the data is perfectly sampled and selected, the bias can also be pre-existing in the data during the generation process. For example, pre-existing bias can occur when the data itself reflects biased decisions, which usually leads to the system being no longer objective and fair. Suppose a company recruiter tends to hire employees with certain race, and the company wants to train a hiring decision system based on its previous recruitment records. The learned system will have a systematic favour towards certain races of people when making hiring decisions. In this case, the AI system will reinforce the pre-existing bias encoded by human decision makers.

2.1.2 ***Algorithmic Bias****.* Moreover, unfairness can arise from the biases in algorithms themselves, such as the improper use of certain optimization methods or biased estimators [12]. Here we list two representative algorithmic biases which may bring unfairness and affect user satisfaction.

Presentation Bias [12]. Presentation bias happens when information is presented in biased ways. For example, ranking bias is a common issue in ranking problems, where the top-ranked results are usually considered as the most relevant results by users and thus will get more clicks, while the lower-ranked results will get fewer exposure even if they are also very relevant. In recommender systems, a well-known issue is popularity bias where the items with more interactions will be recommended more frequently and get more exposure than those less popular but equally or even more relevant ones.

Evaluation Bias [170]. The evaluation bias usually occurs when inappropriate benchmarks are used in model evaluation. Examples include Adience and IJB-A benchmarks, which are biased to skin color and gender when evaluating facial recognition systems [129].

The bias may lurk in the data in lots of ways. It is worth noting that the biases are intertwined due to the feedback loop phenomenon [48]. The data used for training machine learning models are usually collected from user behaviors which may present inherent biases, however, the user behaviors can also be affected by the biased learning algorithms, resulting in the further introduction of bias in the data generation process. Therefore, it is important to consider how the biases affect each other to solve them accordingly [129].

2.1.3 ***Other Causes.*** It is important to notice that bias may not be the only reason for unfairness and there could be other causes, as a result, researchers should always take caution. One example is the conflict between different fairness requirements. Researches have shown that some fairness requirements cannot be satisfied simultaneously [47, 103, 149], therefore, the violation of one type of fairness may be caused by ensuring another. We will introduce more details on the relationship between fairness definitions in the following parts of the survey.

2.2 Methods for Fair Machine Learning

Generally, methods that achieve fairness in machine learning fall under three categories: *Pre-processing*, *In-processing*, and *Post-processing* [35, 53, 129, 206].

2.2.1 ***Pre-processing Method.*** As the bias issue is often in data itself, a straightforward way is to pre-process the training data before the learning process to remove the underlying bias. The key idea is to train a model on a "corrected" data so that the model can be naturally fair. The pre-processing methods are usually achieved through changing the label values for certain data points, or mapping the data to a transformed space. The advantages of pre-processing methods are that the transformed data can be used to train any downstream algorithms without certain assumptions. However, it may suffer unpredictable loss in accuracy and may not remove unfairness on the test data. What's more, the pre-processing method can only be applied if modifying the training data is allowed, both in technical and legal senses.

2.2.2 ***In-processing Method.*** The in-processing methods usually try to balance the accuracy and fairness demands through modifying the learning process. Most of the existing works of fairness in machine learning use this way to remove discrimination. The methods are often achieved through incorporating fairness metrics into the objective function of the main learning task. The in-processing methods usually show good performance on both accuracy and fairness and provide higher flexibility to trade-off between them. The main disadvantage is that such strategy usually leads to a non-convex optimization problem and cannot guarantee optimality.

2.2.3 ***Post-processing Method.*** The post-processing methods apply transformations to the model output so as to mitigate unfairness. This is usually achieved by reassigning the labels assigned by the base models, such as recomputing the scores or re-ranking the output lists. The post-processing methods usually treat the base model as a black-box and provide model-agnostic flexibility. Such methods do not need to modify the model and training data, and can achieve relatively good performance and fairness. However, the post-processing methods cannot be used when sensitive feature information is unavailable at the decision time (test-time) since they usually need to access the sensitive information for post-processing.

In short, there are various ways to improve fairness in machine learning, and each method has its own pros and cons. Choosing which method(s) to use is sometimes not a pure technical problem, but the social and legal context must also be considered. In the following sections, we will introduce examples of using pre-processing, in-processing, and post-processing methods for improving fairness.

2.3 Fairness Definition

An accepted fact in fairness study is that there is no consensus on fairness definitions since the fairness demands can be different under different scenarios. The fairness concerns can be put forwarded from various perspectives, and it has been theoretically proven that some fairness requirements cannot be satisfied at the same time [47, 103, 149]. Therefore, it is important to carefully choose the fairness measurements according to specific problems and scenarios, and be mindful of the potential trade-off between fairness demand and model accuracy. Up to now, many fairness definitions have been proposed in the literature. In general, the definitions of fairness can be classified into three categories: *Group Fairness*; *Individual Fairness*; and *Hybrid Fairness* [175]. The measurement and evaluation of fairness may vary depending on its specific definition. We first provide a general description of the three types of fairness definitions here, and the specific examples in different machine learning tasks will be introduced in the following sections.

2.3.1 ***Group Fairness****.* In the study of group fairness requirements, the research subjects are usually divided into different groups based on a certain grouping method, and the most common way is to split users by their sensitive features. There is no once-and-for-all answer regarding which features are considered sensitive features since different people may have different definitions, while the connotation of sensitive features usually refers to the features that human-beings cannot choose freely at birth or throughout their life such as gender, race and age. Group fairness requires that the protected groups should be treated similarly as the advantaged groups by the machine learning systems [34, 35, 129, 175]. For example, to study the unfairness problem of gender discrimination in a hiring decision-making system, the candidates will be first spilt into different groups according to their gender, and then fairness can be measured through computing the difference of system treatments among different groups such as salary or hiring rate.

2.3.2 ***Individual Fairness****.* Different from group fairness, individual fairness requires that similar individuals should be treated similarly [34, 35, 129, 175]. To guarantee individual fairness, we usually first need to determine whether two individuals are similar according to a certain metric, for example, according to the distance of user representations. The similarity of individuals could be determined by certain sensitive features or combination of features. For example, we may consider all male users whose ages fall into the same range as similar individuals. Sometimes, we may also use latent features to determine user similarity, e.g., we may first represent each user as a latent representation vector according to their profiles or interaction records, and then decide similar users based on vector similarity. Still consider the above example which is to study the unfairness problem in a hiring decision-making system. We can first represent each candidate as a vector and then require that those candidates with close representations have similar salaries.

It is worthwhile to note that group fairness and individual fairness are two orthogonal concepts. For example, suppose we put similar individuals into the same group, individual fairness emphasizes that individuals within the same group are treated fairly, while group fairness emphasizes that different groups are treated fairness on aggregated group-level. An individually fair method may not be fair group-wise, since it is possible that people within the same group are treated fairly but some groups are much better than other groups. On the other hand, a group-wise fair method may not be individually fair, since it is possible that the groups are treated fairly in terms of group-aggregated metrics but some individuals in a group are treated much better than others in the same group. As a result, neither group fairness nor individual fairness subsume the other, and they both need to be considered carefully.

2.3.3 ***Hybrid Fairness.*** Instead of focusing on one certain fairness consideration, hybrid fairness recognizes the fact that the fairness demands vary in most cases and aims to achieve more than one fairness requirements simultaneously. For example, in fair recommendation studies, fairness demands may come from both user-side and item-side, thus multiple fairness definitions should be satisfied at the same time [1, 130]. However, it is worth noting that not all fairness definitions can be guaranteed at the same time since they sometimes may essentially contradict with each other [47, 103, 149].

3 FAIRNESS IN CLASSIFICATION

To better understand the concepts introduced above, we briefly talk about fairness in classification as an example since it is a basic and important task in machine learning and has been relatively well-studied in fairness research. Moreover, recommendation problems can sometimes be formulated as a classification task, e.g., when we try to predict whether the user will click an item or not. Pedreshi et al. [146] first explored fair classification to avoid discrimination in classification rule mining. In recent years, a flurry of methods are proposed for promoting fairness in classification. In the following, we first talk about certain fairness concerns and measurements in classification task, and then introduce the methods together with examples for solving the unfairness issue.

3.1 Fairness Concerns

Without losing generality, we mainly introduce the fairness research in binary classification task here. In a binary classification task, there is training data $\mathcal{D}_{\mathcal{T}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ with user feature vectors $\boldsymbol{x} \in \mathbb{R}^d$ and the corresponding class labels $y \in \{-1, 1\}$. The aim of binary classification problem is to predict the label $\hat{y}_i$ through learning a mapping function $f_{\theta}(\boldsymbol{x}_i)$. For example, $\hat{y}_i = 1$ if $f_{\theta}(\boldsymbol{x}_i) > 0.5$ and $\hat{y}_i = -1$ otherwise. In the literature of fairness in binary classification, each user has an associated sensitive feature $s \in \{0, 1\}$. The aim of fair classification is to avoid the unethical interference of sensitive features into the decision-making process. To this end, several fairness notions have been proposed to measure the unfairness of classifiers [35, 175]. The two basic frameworks in recent studies on fair classification are group fairness and individual fairness. More formally, we introduce the representative fairness notions as follows:

3.1.1 ***Group fairness.*** Group fairness requires that the protected groups should be treated similarly as the advantaged groups. Most of the fair classification studies focus on group fairness concerns. The group fairness notions include the *predicted positive rate*-based metrics and the *confusion matrix*-based metrics [35]. *Predicted positive rate*-based metrics require the parity of the predicted positive rates $\Pr(\hat{y} = 1)$ across different groups, while *confusion matrix*-based metrics take the True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), and False Negative Rate (FNR) into consideration so that the differences between groups can be captured on a more detailed granularity [35]. One example of *Predicted positive rate*-based metrics is:

Statistical Parity: The Statistical Parity, also called *Demographic Parity* or *No Disparate Impact*, requires that each group should have the same likelihood to be classified as positive [60, 206]:

$$\Pr(\hat{y} = 1 \mid s = 0) = \Pr(\hat{y} = 1 \mid s = 1)$$

The shortcoming of this notion is that it ignores the difference between groups. For example, the less popular items may be unpopular for a reason, e.g., due to its unsatisfactory quality. As a result, forcefully requiring all of the unpopular items to have the same click through rate as popular items could be unreasonable. Besides the metrics that focus on the variants of the predicted positive rate $\Pr(\hat{y} = 1)$, most of the group fairness criteria are built on confusion matrix. Examples include:

Equal Opportunity: The Equal Opportunity fairness requires that the True Positive Rate (TPR) is the same across different groups [85]:

$$\Pr(\hat{y}=1 \mid y=1, s=1) = \Pr(\hat{y}=1 \mid y=1, s=0)$$

Equalized Odds: Stricter than Equal Opportunity, the Equalized Odds fairness also takes False Positive Rate (FPR) into account and requires that different groups should have the same true positive rate and false positive rate [14]:

$$\begin{aligned} &\Pr(\hat{y}=1 \mid y=1, s=1) = \Pr(\hat{y}=1 \mid y=1, s=0) \\ \& &\Pr(\hat{y}=1 \mid y=-1, s=1) = \Pr(\hat{y}=1 \mid y=-1, s=0) \end{aligned}$$

Overall Accuracy Equality: The Overall Accuracy Equality fairness requires the same accuracy across groups [14]:

$$\begin{aligned} &\Pr(\hat{y}=-1 \mid y=-1, s=1) + \Pr(\hat{y}=1 \mid y=1, s=1) \\ = &\Pr(\hat{y}=-1 \mid y=-1, s=0) + \Pr(\hat{y}=1 \mid y=1, s=0) \end{aligned}$$

Equalizing Disincentives: The Equalizing Disincentives fairness requires the same difference between the True Positive Rate (TPR) and False Positive Rate (FPR) across groups [94]:

$$\begin{aligned} &\Pr(\hat{y}=1 \mid y=1, s=1) - \Pr(\hat{y}=1 \mid y=-1, s=1) \\ = &\Pr(\hat{y}=1 \mid y=1, s=0) - \Pr(\hat{y}=1 \mid y=-1, s=0) \end{aligned}$$

Treatment Equality: The Treatment Equality fairness requires the same ratio of False Negative Rate (FNR) to False Positive Rate (FPR) across groups [14]:

$$\frac{\Pr(\hat{y}=1 \mid y=-1, s=1)}{\Pr(\hat{y}=-1 \mid y=1, s=1)} = \frac{\Pr(\hat{y}=1 \mid y=-1, s=0)}{\Pr(\hat{y}=-1 \mid y=1, s=0)}$$

3.1.2 ***Individual fairness.***

Instead of considering fairness across groups, individual fairness notions require that similar individuals should be treated similarly [35, 175]. Examples include:

Counterfactual Fairness: Counterfactual Fairness is an individual-level causal-based fairness notion [105]. It requires that for any possible individual, the predicted result of the learning system should be the same in the counterfactual world as in the real world. Given a set of latent background variables U, the predictor $\hat{Y}$ is counterfactually fair if under any context $\boldsymbol{X}=\boldsymbol{x}$ and $S=s$, the equation below holds for all y and for any value s' attainable by S:

$$\Pr\left(\hat{Y}_{S\leftarrow s}(U)=y \mid \boldsymbol{X}=\boldsymbol{x}, S=s\right) = \Pr\left(\hat{Y}_{S\leftarrow s'}(U)=y \mid \boldsymbol{X}=\boldsymbol{x}, S=s\right)$$

It is worth noting that counterfactual fairness is a notion defined based on the idea of causality which emphasizes the independence between the sensitive features and the predicted outcomes. However, the techniques to achieve counterfactual fairness can be diverse and are not limited to typical causal inference methods such as interventions, examples include: simply removing the sensitive features and their descendants from the model and prediction function [105]; variational autoencoders [45]; adversarial learning [83]; data pre-processing [41]; causal regularization [55]; data augmentation [124], etc.

Fairness Through Awareness: The Fairness Through Awareness requires that any two individuals with similar non-sensitive features should receive similar predicted results [60]. Let's consider two individuals $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The distance between the two individuals is defined by $d(\boldsymbol{x}_1, \boldsymbol{x}_2)$, and the difference of predicted results between the two individuals is computed through $F(\hat{y}_1, \hat{y}_2)$. Here, the formulation of $F(\cdot,\cdot)$ and $d(\cdot,\cdot)$ are usually determined by specific tasks. The Fairness Through Awareness requires that:

$$F(\hat{y}_1, \hat{y}_2) \le \alpha \cdot d(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

which means that the difference in treated should be upper bounded by the difference between the two individuals.

3.2 Methods

In this section, we introduce the methods for promoting fairness in classification. As we talked above, the techniques of fair classification can mainly be divided into three categories: pre-processing, in-processing and post-processing [206]. The methods usually first specify one or more fairness notions to achieve and then propose corresponding methods to control the selected notions.

3.2.1 ***Pre-processing Method.*** The pre-processing methods focus on mitigating the bias in data so that any model learned from the data will be fair [206]. The pre-processing methods can be applied through changing the labels for certain subjects [95, 123], mapping the data to a transformed space [33], or perturbing the non-sensitive features [68]. For example, Calders et al. [32] study the problem of unfairness in classification and consider the case where there is unjustified dependency between sensitive features and class labels in the input data. They propose two methods to pre-process the training data with the goal of reducing the dependency between sensitive features and class labels while maintaining the overall positive class probability: massaging and re-weighting. For the massaging method, they change the labels of some subjects x with sensitive features $S = s$ from "−" to "+", and meanwhile change the labels of the same number of objects with $S \neq s$ from "+" to "−". For the re-weighting method, they re-weight the subjects to reduce the dependency. For example, subjects with $S = s$ and class label "+" will get higher weights than subjects with $S = s$ and class label "−". Subjects with $S \neq s$ and class label "+" will get lower weights than subjects with $S \neq s$ and class label "−". The authors conduct experiments to show that both of the methods can remove the dependency from training data so that the classifier learned from the data would be fairer. However, the accuracy could be sacrificed after the pre-processing.

3.2.2 ***In-processing Method.*** Most of the works adopt the in-processing methods to modify the training procedure of the classifier [79, 97, 152, 183]. For example, Zafar et al. [206] introduce a flexible constraint-based framework to enable the design of fair margin-based classifiers. To design a fair convex boundary-based classifier, the authors propose to minimize the corresponding loss function under fairness constraints during training:

$$\left.\begin{array}{rl} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & P(. \mid s = 0) = P(. \mid s = 1) \end{array}\right\} \text{Classifier loss function} \tag{1}$$

Here, the fairness constraints can be replaced with any fairness requirement that we want. Take statistical parity as an example, it requires the following constraint:

$$\Pr(\hat{y} = 1 \mid s = 0) = \Pr(\hat{y} = 1 \mid s = 1) \tag{2}$$

In particular, the authors consider to control the covariance $\mathrm{Cov}_{SP}(s, d_{\boldsymbol{\theta}}(\boldsymbol{x}))$ between the users' sensitive attribute s and the signed distance $d_{\boldsymbol{\theta}}(\boldsymbol{x})$ from the users' feature vector $\boldsymbol{x}$ to the decision boundary, since this distance decides the value of the predicted label $\hat{y}$.

$$\mathrm{Cov}_{SP}(s, d_{\boldsymbol{\theta}}(\boldsymbol{x})) = \mathbb{E}\left[(s - \bar{s}) d_{\boldsymbol{\theta}}(\boldsymbol{x})\right] - \mathbb{E}[(s - \bar{s})] \bar{d}_{\boldsymbol{\theta}}(\boldsymbol{x}) \approx \frac{1}{N} \sum_{(\boldsymbol{x}, s) \in \mathcal{D}_{\mathcal{T}}} (s - \bar{s}) d_{\boldsymbol{\theta}}(\boldsymbol{x}) \tag{3}$$

To satisfy statistical parity, we need to make sure that the sensitive feature s is irrelevant to the distance $d_{\boldsymbol{\theta}}(\boldsymbol{x})$, and thus the empirical covariance defined above needs to be approximately zero.

Therefore, the we can finally get the objective function as bellow:

$$\begin{array}{ll} \text{minimize} & L(\boldsymbol{\theta}) \\ \text{subject to} & \frac{1}{N}\sum_{(\boldsymbol{x},s)\in\mathcal{D}_{\mathcal{T}}}(s-\bar{s})d_{\boldsymbol{\theta}}(\boldsymbol{x}) \leq c \\ & \frac{1}{N}\sum_{(\boldsymbol{x},s)\in\mathcal{D}_{\mathcal{T}}}(s-\bar{s})d_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq -c \end{array} \tag{4}$$

Here, $c \in \mathbb{R}^{+}$is a given threshold to trade off the accuracy and unfairness due to statistical parity. The above optimization problem is convex and the trade-off between the classifier accuracy and fairness constraint is Pareto-optimal. The conducted experiments which simulate statistical parity in classification outcomes show that the fairness constraints could improve fairness, but would suffer sacrifice on accuracy [206].

3.2.3 ***Post-processing Method.*** Post-processing methods focus on modifying the scores learned from the base classifiers so that the new results are fair. For example, the post-processing methods can mitigate unfairness through learning a different decision threshold for a given scoring function or re-weighting the features [49, 61, 85, 132, 149]. In Mehrabi et al. [128], the authors design a post-processing bias mitigation strategy based on attention. The proposed method is flexible and can be used for any fairness notion. First, an attention-based classifier is trained and we can get the attention weight of each input feature. To find out the effect of each feature on the fairness of outcomes, the authors zero out or reduce the attention weight of each feature and measure the difference in fairness of the outcomes based on the desired fairness notion such as Equalized Odds. If the difference on classification accuracy is small but the difference on fairness is large, it indicates that this feature is mostly responsible for unfairness, and it may need to be dropped to mitigate unfairness through reducing its attention weight. The authors can also control the fairness-accuracy trade-off by decreasing the weights of the features that will hurt fairness or accuracy, and increasing the attention weights of the features that can boost accuracy while keeping the fairness.

As for the evaluation of fairness, the metrics usually highly depend on the specific fairness definition or requirement. For example, if we consider statistical parity, we can adopt $SP = |P(\hat{y} = 1 \mid s = 0) - P(\hat{y} = 1 \mid s = 1)|$ to evaluate fairness. The datasets for studying fairness in classification usually contain sensitive features, which have been collected and summarized in several previous surveys such as [129, 153]. Therefore, we would not repeat the introduction here, but we will present several datasets for fair recommendation tasks in Section 5.5.

4 FAIRNESS IN RANKING

Besides fairness studies in classification task, recent works have also raised the question of fairness in rankings. Recommendation algorithms can usually be considered as a type of ranking problem. Therefore, the fairness studies in ranking is pretty enlightening to improve fairness in recommendations. However, the existing works of fair ranking usually only consider unfairness issue from the item side, i.e., the candidates to be ranked, while the concept of fairness in recommender systems has been extended to multiple stakeholders [29]. In this section, we aim to briefly introduce the research on fair ranking and then focus on fairness in recommendation beginning from the next section.

4.1 Fairness Concerns

There are two main types of ranking task: *Score-based Ranking* and *Learning to Rank* [209]. The difference between score-based ranking and learning to rank is how to obtain the score for ranking. In score-based ranking, the score can be directly computed from a given function, while in learning to rank, the score is estimated by training a model from the preference-enriched training examples.

To discuss the fairness works from a machine learning view, we focus on talking about the fairness studies in learning to rank problem.

In the supervised learning to rank task, a set of candidates is given as $\{c_1, c_2, \ldots c_n\}$, where each candidate is described by a set of features X, which may include sensitive features S. The goal of a ranking algorithm is to generate a ranking list l which sorts the candidates by their predicted relevance to the search query. The highest-scoring candidates will appear closer to the top of the list and get the most exposure. We typically return the best-ranked K candidates to construct a top-K ranking list. A majority of the existing works on fair ranking focus on list-wise definitions for fairness, which means that the fairness measures depend on the entire list of results for a given query. As we mentioned above, there is also no universal fairness measure in ranking tasks due to the complex system environment and goals. The fairness notions proposed in classification tasks may be adopted in ranking tasks with some additional task-specific considerations. Generally, the fairness definitions in ranking can also be classified into group and individual fairness. However, in this survey, we talk about the taxonomy of fairness in ranking from a more task-specific perspective: *Probability-based Fairness* and *Exposure/Attention-based Fairness* [144, 209].

4.1.1 ***Probability-based Fairness.*** The probability-based fairness notions in ranking usually require a minimum/maximum number or proportion of the protected candidates in the top-K ranking list [11, 36, 77, 207]. For example, in paper [36], the fairness is required as:

$$L_{k\ell} \leq \sum_{1 \leq j \leq k} \sum_{i \in P_\ell} x_{ij} \leq U_{k\ell} \tag{5}$$

where x is a binary assignment matrix and $x_{ij} = 1$ if item i is assigned to position j. The upper and lower bound $U_{k\ell}, L_{k\ell} \in \mathbb{Z}_{\geq 0}$ guarantees that a certain number of items with property ℓ appear in the top-K positions of the ranking. It is worth mentioning that probability-based fairness definitions are proposed for achieving group fairness goals.

4.1.2 ***Exposure/Attention-based Fairness.*** Another type of fairness definition in ranking is Exposure or Attention-based Fairness [56, 135, 167, 208]. In fair ranking task, a frequently studied problem is how to distribute the exposure opportunity to candidates fairly since the total exposure of candidates is a limited resource. The competition of exposure comes from the well-known issue in ranking: the position bias [51], which means that the expected exposure of candidates or attention from users will reduce significantly as the ranking position increases, and thus a slight difference in estimated relevance could result in a large difference in item exposures. In such setting, the fairness metrics are usually relevant to the exposure of the candidates belonging to different groups, so that the average exposure of groups can be controlled to be proportional to their average relevance to the search query. Different from probability-based fairness notions, the exposure/attention-based fairness can quantify not only group fairness [135, 167], but also individual fairness [18, 26] according to specific formalization. Group fairness can be measured through the difference of the average exposure between different groups G_1 and G_2 [209]:

$$F(G_1, G_2) = \left| \frac{1}{|G_1|} \sum_{c \in G_1} \text{Exposure}(c) - \frac{1}{|G_2|} \sum_{c \in G_2} \text{Exposure}(c) \right|$$

Similarly, individual unfairness can be measured through the difference of exposure between two candidates c_1 and c_2 [209]:

$$F(c_1, c_2) = |\text{Exposure}(c_1) - \text{Exposure}(c_2)|$$

4.2 Methods

In this section, we introduce the methods for achieving fairness in ranking. Similarly, we also follow the order of pre-processing, in-processing and post-processing. The majority of works of fair ranking adopt in-processing and post-processing methods. The advantages and disadvantages of the three types in ranking are similar as we discussed in Section 2.2. In this section, we follow the notations of each paper and explain the notions so that readers can easily reference and understand them.

4.2.1 ***Pre-processing Method.*** Pre-processing methods seek to mitigate bias in training data, so that the models trained from the unbiased data will be fair [106, 169]. For example, in Lahoti et al. [106], authors introduce a method for mapping user records into a low-rank representation that reconciles individual fairness through pre-processing. Specifically, given two samples x_i and x_j, and let's use x_i^* and x_j^* to denote their representations with only non-sensitive features. The method aims to learn a mapping function ϕ so that the individuals who are indistinguishable on their non-sensitive features x^* in data should also be indistinguishable in their transformed representations $\tilde{x}$ under a given distance function d:

$$\left|d\left(\phi\left(x_i\right), \phi\left(x_j\right)\right) - d\left(x_i^*, x_j^*\right)\right| \leq \epsilon$$

To this end, the problem is formalized as a probabilistic clustering problem. Specifically, we aim for K clusters, each given in the form of a prototype vector v_k with $k = 1 \cdots K$. Every sample x_i is assigned to a cluster according to a probability distribution u_i which reflects the distance of the sample x_i from prototypes. Here the u_{ik} represents the probability of x_i belonging to the cluster of prototype v_k. Therefore, the representation $\tilde{x}_i$ is given by:

$$\tilde{x}_i = \phi\left(x_i\right) = \sum_{k=1\cdots K} u_{ik} \cdot v_k$$

To ensure the utility of any downstream application, we can optimize the utility objective through minimizing the data loss induced by ϕ:

$$L_{util}(X, \tilde{X}) = \sum_{i=1}^{M} \|x_i - \tilde{x}_i\|_2 = \sum_{i=1}^{M} \sum_{j=1}^{N} \left(x_{ij} - \tilde{x}_{ij}\right)^2$$

Furthermore, to reduce individual unfairness in data, we also need to optimize the fairness objective as:

$$L_{fair}(X, \tilde{X}) = \sum_{i,j=1\ldots M} \left(d\left(\tilde{x}_i, \tilde{x}_j\right) - d\left(x_i^*, x_j^*\right)\right)^2$$

The final objective combines the utility loss and fairness loss above as:

$$L = \lambda \cdot L_{util}(X, \tilde{X}) + \mu \cdot L_{fair}(X, \tilde{X})$$

where λ and μ are hyper-parameters, and the objective can be optimized through gradient descent. The transformed representation reconciles individual fairness and can be incorporated into a variety of tasks such as classification and learning-to-rank. Experiments on learning-to-rank show that the proposed methods achieve best individual fairness but worse utility than baseline models [106].

4.2.2 ***In-processing Method.*** Most of the works on fair ranking adopt in-processing methods, which aim to directly learn a fair ranking model from scratch [139, 168, 208]. For example, Narasimhan et al. [139] study the problem of learning pairwise fairness for ranking. Suppose there are queries S drawn *i.i.d.* from an underlying distribution D, each candidate to be ranked for answering the query is associated with a vector $x \in X$ and a label $y \in Y$. For example, Y may take the value of 1 or 0 to represent whether the result is clicked by a user. The ranking algorithm needs to learn a scoring function $f : X \longrightarrow \mathbb{R}$ for sorting the candidates. Suppose there are a set of K groups $G_1, \cdots, G_K$ and each example belongs to exactly one group. The authors define the group-dependent pairwise accuracy $A_{G_i > G_j}$ as the probability of a clicked candidate from group G_i being ranked above another relevant unclicked candidate from group G_j:

$$A_{G_i > G_j} := P\left(f(x) > f\left(x'\right) \mid y > y', (x, y) \in G_i, \left(x', y'\right) \in G_j\right) \tag{6}$$

where (x, y) and (x', y') are drawn *i.i.d.* from the distribution of examples. The cross-group pairwise equal opportunity is defined as:

$$A_{G_i > G_j} = \kappa, \text{ for some } \kappa \in [0, 1] \text{ for all } i \neq j. \tag{7}$$

The training problem can be formulated as maximizing the overall pairwise accuracy under the cross-group equal opportunity fairness constraint:

$$\begin{aligned} &\max_{f \in \mathcal{F}} \mathrm{AUC}(f) \\ \text{s.t.} \quad &A_{G_i > G_j}(f) - A_{G_k > G_l}(f) \leq \epsilon \quad \forall i \neq j, k \neq l. \end{aligned} \tag{8}$$

where $\mathrm{AUC}(f)$ is the overall pairwise accuracy and $\mathcal{F}$ is the class of models we are interested in. Experiment results show that the proposed method can greatly reduce the fairness violation at the cost of a lower test AUC. The paper claims that their algorithm can be applied to any pairwise metric.

4.2.3 ***Post-processing Method.*** Post-processing methods usually re-rank candidates in the output list of the base ranking models [18, 119, 207]. For example, Singh and Joachims [167] study the problem of fair exposure in rankings. They consider the setting where the relevance of documents has been obtained or well estimated from some base rankers, and they only ask how to fairly allocate exposure in rankings based on the known relevance. For a given single query q, the ranking algorithm aims to present a ranking r of a set of documents $\mathcal{D} = \{d_1, d_2, d_3 \cdots d_N\}$. The utility of a ranking r for query q is denoted as $\mathrm{U}(r|q)$. The authors propose to optimize the ranking task under fairness constraints so that the learned ranking r maximizes the utility function and fairness. To learn the distribution of the ranking r, the authors learn a probabilistic ranking matrix P where $\mathbf{P}_{i,j}$ is the probability that r places document d_i at rank j. Here, we need both the sum of probabilities for each position and the sum of probabilities for each document to be 1 , i.e. $\sum_i \mathbf{P}_{i,j} = 1$ and $\sum_j \mathbf{P}_{i,j} = 1$. Therefore, the problem of optimal fair ranking can be formulated as finding the utility-maximizing probabilistic ranking P under fairness constraints:

$$\begin{aligned} \mathbf{P} = \operatorname{argmax}_{\mathbf{P}} \quad & \mathrm{U}(r|q) && \text{(expected utility)} \\ \text{s.t.} \quad & \mathbb{1}^T \mathbf{P} = \mathbb{1}^T && \text{(sum of probabilities for each position)} \\ & \mathbf{P}\mathbb{1} = \mathbb{1} && \text{(sum of probabilities for each document)} \\ & 0 \leq \mathbf{P}_{i,j} \leq 1 && \text{(valid probability)} \\ & \mathbf{P} \text{ is fair} && \text{(fairness constraints)} \end{aligned} \tag{9}$$

The proposed method is able to achieve various fairness notions in ranking depending on specific fairness requirements. Here we also take Statistical Parity of exposure as an example. First of all, the exposure for a document d_i under a probabilistic ranking P is defined as:

$$\text{Exposure } (d_i|\mathbf{P}) = \sum_{j=1}^{N} \mathbf{P}_{i,j} \mathbf{v}_j \tag{10}$$

where v_j represents the importance of position j and the exposure for a group G_k is the average exposure of each document in G_k:

$$\text{Exposure}\left(G_k|\mathbf{P}\right) = \frac{1}{|G_k|} \sum_{d_i \in G_k} \text{Exposure}\left(d_i|\mathbf{P}\right) \tag{11}$$

Based on this, the Statistical Parity constraint requires that:

$$\text{Exposure}\left(G_0|\mathbf{P}\right) = \text{Exposure}\left(G_1|\mathbf{P}\right) \tag{12}$$

After derivation and simplification in this paper, the fairness constraint can be plugged into the linear program to solve. The experiments show that the effect of statistical parity in ranking is similar to its effect in classification, where it can also lead to a drop of accuracy.

Similarly, the evaluation of fairness in ranking is also highly relevant to the specific fairness definition or requirement. The survey [209] summarized some datasets for studying fairness in ranking.

5 FAIRNESS IN RECOMMENDATION

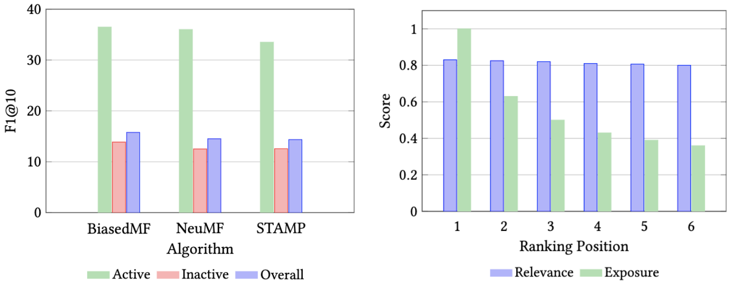
In this section, we introduce the definitions, methods, evaluation, datasets and challenges for fairness in recommendation research. We first briefly talk about the preliminaries of recommender systems and show several examples of unfairness in recommendation. After that, we systematically show the taxonomy of fairness notions, techniques of promoting fairness, evaluating fairness, and datasets for studying fairness in recommendation. We also analyze the challenges and opportunities in fair recommendation field to inspire more works in the future.

5.1 Preliminaries of Recommender Systems

In recommendation task, there is usually a user set $\mathbb{U} = \{u_1, u_2, \cdots, u_n\}$ and an item set $\mathbb{V} = \{v_1, v_2, \cdots, v_m\}$, where n is the number of users and m is the number of items. We also have the user-item interaction histories represented as a 0-1 binary matrix $H = \left[h_{ij}\right]_{n \times m}$, where each entry $h_{ij} = 1$ if user u_i has interacted with item v_j, otherwise $h_{ij} = 0$. The main task for recommendation is to predict the preference scores of users over items S_{uv}, so that the model can recommend each user u_i a top-N recommendation list $\{v_1, v_2, \cdots, v_N|u_i\}$ according to the predicted scores. To learn preference scores, modern recommender models usually learn user and item representations from data, and then take the representations as input to a learned or designed scoring functions to make recommendations. Most of the Collaborative Filtering (CF) [63, 80, 160] or Collaborative Reasoning (CR) [40, 42, 164] recommender systems are directly trained from user-item interaction history. The content-based recommendation models [7, 121, 215] and hybrid models [28] may also leverage user/item profiles as input or use additional information to help train the model. In fair recommendation works, fairness concerns are usually relevant to user and item features such as user gender or race, and item category or popularity.

5.2 Examples of Unfairness in Recommendation

The fairness considerations can be raised from very different perspectives in recommendation task. We show two examples of unfairness in recommendation in Fig. 1, one from user-side and another from item-side. Fig.(1a) shows the unfairness of recommendation quality between groups of active users and inactive users in an e-commerce dataset [112]. We first sort users according to their number of interactions in the training data. We then label the top 5% users as active ones and leave the remaining 95% users into the inactive group. After that, we test the performance of three fairness-unaware recommendation models (BiasedMF [104], a shallow model, NeuMF [86], a deep model, and STAMP [118], a sequential model) on two user groups. We can see that



(a) User-side unfairness

(b) Item-side unfairness

Fig. 1. (a) The significant difference between inactive and active user groups on recommendation quality, image from [112]; (b) The significant exposure difference among items with close relevance scores.

the recommendation quality (F1 score) of active users is significantly higher than those inactive ones, which shows that the recommendation models are dominated by the active users and thus generate unfairness treatments for inactive users, which also degrade the overall performance since the inactive users are the majority. Similar phenomenon is also observed on some other datasets [23, 154]. In addition, Fig.(1b) shows the unfairness of item exposure in recommendation lists. Suppose there are a set of items being recommended to a user and each item has a learned relevance score. The recommender systems usually sort the items by their relevance scores and the highly scored items will appear closer to the top of the recommendation list. Here we consider a standard exposure drop-off which is commonly used in the Discounted Cumulative Gain measure: $1/\log(1+i)$ to compute the exposure of each position, where i is the position in the list [167]. We can see from the figure that although the items have very close relevance scores, the exposure opportunity received by each item significantly varies, which causes unfairness of exposure for items in recommendation. A similar example is also discussed in [167].

5.3 A Taxonomy of Fairness Notions in Recommendation

In this section, we introduce the taxonomy of fairness notions in recommendation to provide a systematic overview of the various perspectives of fairness in recommendation [115, 116]. In Table.1, we summarize the categories as well as some papers in each category.

5.3.1 ***Group Fairness vs. Individual Fairness.*** As we introduced above, group fairness and individual fairness are two basic frameworks in algorithmic fairness and the fairness in recommendation can also be divided into these two categories. Group fairness requires that the protected group should be treated similarly to the advantaged group or the population as a whole, while individual fairness demands that similar treatment should be received by each similar individual. Here we introduce some representative examples for each category. Yao and Huang [202] present four group fairness metrics in collaborative filtering recommender systems. The authors consider a binary group feature such as gender to divide disadvantaged and advantaged groups and all proposed metrics measure a discrepancy between the prediction behavior for disadvantaged users and advantaged users. Li et al. [112] consider user-oriented group fairness in commercial recommendation. The authors divide users into advantaged and disadvantaged groups according to their activity, e.g., the number of user interactions in the training data, and require that a fair recommendation algorithm should offer similar recommendation quality for different groups of users. The authors use the difference of the average recommendation performance (such as F1 and NDCG) between two groups to measure the user group unfairness of a recommendation algorithm.

Table 1. The Taxonomy of Fairness Notions in Recommender Systems.

Criterion	Category	Reference
Level	Group Individual	[219] [130] [221] [77] [138] [112] [73] [92] [81] [173] [137] [75] [142] [127] [114] [108] [57]
Subject	User Item	[107] [70] [100] [112] [114] [92] [189] [191] [161] [57] [185] [184] [130] [15] [77] [138] [73] [220] [81] [163] [178] [108] [151] [201]
Side	Single-side Multi-side	[107] [70] [220] [81] [112] [114] [108] [92] [189] [186] [184] [151] [31] [1] [142] [143] [190] [187] [19] [134] [188] [137]
Relationship	Associative Causal	[202] [67] [219] [38] [20] [108] [22] [220] [184] [155] [39] [112] [114] [89] [177]
State	Static Dynamic	[202] [117] [157] [107] [15] [70] [39] [22] [112] [114] [150] [58] [73] [120] [9]
Duration	Short-term Long-term	[202] [117] [157] [107] [99] [15] [4] [70] [5] [112] [114] [173] [24] [73] [19] [120] [66] [9] [75]
Granularity	Populational Personalized	[24] [70] [112] [73] [220] [120] [81] [184] [151] [188] [137] [9] [114] [191]
Transparency	Blackbox Explainable	[202] [67] [114] [89] [66] [112] [220] [190] [187] [19] [9] [75] [74]

Results show that fairness-aware learning can achieve both better fairness on recommendation quality between the two groups and better overall recommendation performance of the system, which means that when properly treated, fairness and utility may not conflict with each other and they can sometimes be improved simultaneously. Similar results are also observes on more datasets and recommendation scenarios [154]. Furthermore, Rahmani et al. [154] show that the recommendation accuracy disparity among user groups may vary on different datasets. In general, such disparity is more significant in recommendation scenarios where user behaviors require significant cost, such as e-commerce where users really have to pay money to purchase a product and POI recommendation where users have to spend travel costs (both time and money) to check in a location. In these scenarios, the behavioral difference between advantaged and disadvantaged users can be more apparent due to their different ability to afford the costs. However, such disparity can be less severe in some other recommendation scenarios such as music recommendation, since once users have the membership, the cost of listening a song is relatively small (just a few minutes of time). This shows that it is important to carefully define the user groups according to the recommendation scenario when considering group fairness in recommendation.

Some other works consider individual fairness in recommendation. For example, Lin et al. [117] study the problem of group recommendation where items are recommended to groups of users whose preferences can be different from each other. The authors define the individual utility given a recommendation to the group by considering the relevance of each recommended item to the user, and treat the imbalances between group members' utilities as the metric of fairness. Authors maximize the satisfaction of each group member while minimize the unfairness between them. Li et al. [114] consider counterfactual fairness in recommendation, which requires that the recommendation results for each user are the same in the factual and the counterfactual world.

The counterfactual world is defined as the one where user's sensitive features were changed while all the other features that are not causally-dependent on the sensitive features remain the same.

5.3.2 ***User Fairness vs. Item Fairness.*** As a multi-stakeholder system, fairness requirements in recommender systems may come from different sides including but not limited to user-side and item-side, where the user-side refers to the users who receive recommendations and item-side refers to the items to be ranked or recommended. In some cases such as job applicant recommendation for recruiters, the job seekers are those to be ranked and recommended, but we still consider the fairness demands of them as item-side fairness for simplicity, since even though the job applicants are human users of the job matching system, they are the "items" to be ranked from the recommender system's point of view. The fairness demands from user side are usually about the quality of the recommendations for them, while the fairness considerations from item side usually focus on the exposure opportunity of items in the ranking lists. Here we discuss some examples in the following. Fu et al. [70] considers mitigating the unfairness problem for users in the context of explainable recommendation over knowledge graphs. The authors find that performance disparity exists between active and inactive user groups and claim that such disparity may come from the different distribution of path diversity. Wu et al. [184] investigate the problem that big recommendation models are unfair to cold-start users. Specifically, the authors observe that the recommendation accuracy of cold-start users is sacrificed during the process of optimizing the overall performance of big recommendation models. The authors proposed a self-distillation approach to encourage the model to fairly capture the interest distributions of both cold-start and active users. Compared with the works considering user-side fairness, some other research focus on item-side fairness. One typical problem of item fairness in recommendation is to mitigate the exposure unfairness which is usually due to popularity bias, i.e., the popular items may get disproportionately more exposure opportunity since they dominate the embedding learning in recommender systems, while the less popular items may get much less exposure opportunity even though they have equally good or even better quality. The exposure unfairness issue is often solved by increasing the number of unpopular items (long-tail items) or otherwise the overall catalog coverage in the recommendation list, or making sure that the exposure rate of items is proportional to their quality [2, 3, 6, 98]. Another example of item-side fairness in recommendation is [15] which proposes the notion of pairwise fairness in recommendation. The authors measure item fairness based on pairwise comparisons, and require that the likelihood of a clicked item being ranked above another relevant unclicked item should be the same across different item groups, conditioned on that the items have been engaged with the same amount.

5.3.3 ***Single-side Fairness vs. Multi-side Fairness.*** Since the concept of fairness in recommender systems has been extended to multiple stakeholders [29], it is not satisfactory to meet the fairness demands of only user side or item side, and thus an important problem is how to balance the fairness demands of multiple sides. Therefore, besides the single-side fairness considerations as we mentioned above, several works study multi-side fairness in recommender systems. Both [29] and [1] present several most commonly observed classes of multi-stakeholder recommendation, and categorize different types of fairness that are important to address in a multi-stakeholder recommendation environment. Patro et al. [142] address individual fairness for both producers and customers. From the item-side, authors require to reduce the exposure inequality among items, and from the user-side, authors argue that the platforms should fairly distribute the loss in utility among all the customers. Specifically, the authors formulate the problem as an allocation problem that guarantees minimum exposure for the items and envy-free up to one item (EF-1) [27] for the users. Here the EF-1 ensures that every user values his or her allocation at least as much as any other agent's allocation after hypothetically removing the most valuable item from the other agent's allocated bundle. Patro et al. [143] study the scenario where the online recommendation algorithms are updated frequently to improve user utility. The authors argue for incremental updates of the platform algorithms to avoid the abrupt change of item exposures. To ensure the fairness for two-sided platforms, the authors propose an online optimization method to achieve smooth transition of the item exposures while guaranteeing a minimum utility for every customer. From the item side, the authors require a minimal difference between the exposure distribution in the new system and that in the old system, while for the user side, the authors define a platform to be fair for the customers if it guarantees a minimum utility for everyone.

5.3.4 ***Associative Fairness vs. Causal Fairness.*** The research community first studied fairness in machine learning by developing association-based (or correlation-based) fairness notions, with the aim to find the discrepancy of statistical metrics between individuals or sub-populations. Up to now, most of the existing works about fairness in recommendations consider the association-based fairness notions. For example, the previously mentioned research [202], which proposes metrics to measure the discrepancy between the prediction behavior for disadvantaged users and advantaged users in collaborative filtering recommender systems. Recently, researchers have found that fairness cannot be well assessed only based on association notion [101, 105, 211, 212], since they cannot reason about the causal relations among features. However, unfair treatments usually result from a causal relation between the sensitive features (e.g. gender) and model decisions (e.g. admission). Therefore, researchers have proposed causal-based fairness notions [105, 194] to study unfairness issues in machine learning more properly. Unlike the association-based notions which are only computed based on data, the causal-based fairness notions also consider the additional structural knowledge of the system regarding how variables propagate on a causal model (e.g., a causal graph) [125]. The causal-based fairness notions are usually defined in terms of interventions and counterfactuals [145]. The non-observable properties of interventions and counterfactuals bring great challenges to the applicability of causal-based notions in real world since sometimes they cannot be easily computed from observational data [125]. Two main frameworks are proposed to solve the problem: one is known as the potential outcome framework [91], which usually estimates the causal quantities through re-weighting and matching techniques; another is known as the structural causal model [145], which tells how to estimate causal quantities through observable data under identifiability criterion [166]. Li et al. [114] aim to achieve counterfactual fairness in recommender systems. The authors define counterfactually fair recommendation as the recommendation results that are the same in factual and counterfactual world for each possible user. In the paper, the counterfactual world could be the one where user's sensitive features are changed, for example, the gender of a user is changed from male to female, while all other insensitive features which are not causally-dependent on sensitive features remain the same. Compared with the classification problem, causality in fair recommendation has been rarely studied. Since it is now widely accepted that causal-based notions are important to be considered to enhance fairness [125], we believe causality considerations will open up new challenges and opportunities for studying fairness in recommendation.

5.3.5 ***Static Fairness vs. Dynamic Fairness.*** Most of the existing works about fairness in recommendation consider static fairness where the recommendation environment is fixed during the recommendation process, and usually provide a one-time fairness solution based on fairness-constrained optimization. For example, Li et al. [112] divide users into advantaged and disadvantaged groups based on their activity and propose a re-ranking method to mitigate the performance disparity between the two groups. The user activity level is assumed to be unchanged and fixed during the recommendation process. However, recommender systems are usually dynamic systems since users constantly interact with items, as a result, a previously inactive user may now become active while a previously unpopular item may now become popular, and vice versa [126]. What's more, researches have shown that imposing static fairness criteria myopically at every step may actually exacerbate unfairness [52, 54, 182, 214]. To solve the problem, a few works have paid attention to the dynamic factors in the recommender system environment and study how to enhance fairness with dynamics such as the change of utility, attributes and group labels due to the user interactions throughout the recommendation process [210]. Ge et al. [73] study the dynamic fairness of item exposure in recommender systems. The items are separated into advantaged and disadvantaged groups based on item popularity. However, the item popularity may change during the recommendation process based on the recommendation strategy and user feedback, causing the underlying group labels to change over time. To solve the challenge, the authors formulate the problem as a Constrained Markov Decision Process (CMDP) by dynamically constraining the fairness of item exposure at each iteration. Liu et al. [120] study the task of Interactive Recommender Systems (IRS) where items are recommended to users consecutively and the user feedback is received during the process. IRS gradually refine the recommendation policy according to the obtained user feedback in an online manner and aim to maximize the total utility over the whole interaction period. During the recommendation process, the user preferences and the system's fairness status constantly change over time. To resolve this problem, the authors propose a reinforcement learning based framework, which jointly represents the user preferences and the system's fairness status into the states of the Markov Decision Process (MDP) for recommendation, so that they can dynamically achieve a long-term balance between accuracy and fairness in IRS.

5.3.6 ***Short-term Fairness vs. Long-term Fairness.*** The fairness of recommender systems can also be considered from short-term or long-term fairness requirements. Short-term fairness considers to achieve fair recommendations at present, while long-term fairness aims to ensure the fairness in the long run. The short-term fairness is usually a static fairness since it only needs to consider the current status and achieve fairness for one-time. The dynamic fairness can be seen as a type of long-term fairness since the dynamics usually occur in the long run. However, the long-term fairness covers a broader scope such as the case where there is no dynamics but the fairness cannot be realized in the present and can only be addressed by a long term strategy. For example, Borges and Stefanidis [24] consider to achieve the individual exposure fairness for items. In recommendation task, items are presented as an ordered list. For items with very close or equal relevance scores, the algorithm needs to arrange them in a proper order. However, users usually pay more attention to the top positions of the list and the level of attention significantly decreases as the position in the ranking gets lower. In such case, individual exposure fairness (i.e., items with close relevance scores get similar amounts of exposure) cannot be achieved within just one search result due to the limited positions and can only be achieved in the long term by changing the position of items in multiple rounds of ranking so as to achieve fairness in expectation. Specifically, the authors require that the ranked items receive exposure that is proportional to their relevance scores in a series of rankings through an amortized manner.

5.3.7 ***Populational Fairness vs. Personalized Fairness.*** Most fairness definitions consider fairness on populational level which aim to achieve the same fairness definition for all users. However, users' fairness demands can be very personalized. For example, some users may be very sensitive on gender and they want to be fairly treated on the gender feature, while some other users may not care about gender too much but instead they are very sensitive to age and they do not want to be discriminated by age. It is important to understand that populational vs. personalized fairness is different from group vs. individual fairness. The group vs. individual fairness dimension emphasizes whether users are treated as a group or individually, while populational vs. personalized fairness dimension emphasizes whether users have the right to tell the system what fairness consideration that they care about rather than the system designer care about. More specifically, a system guarantees individual fairness does not necessarily mean that it provides personalized fairness—it could be populational. For example, the system may globally apply the same individual fairness such as counterfactual gender fairness to all users, even though some users do not care about gender fairness but care about age fairness instead. Similarity, a system guarantees group fairness does not necessarily mean that it is populational fairness—it could be personalized. For example, the system may split users into two groups and guarantee that the average user satisfaction of group one is fair compared to that of group two, meanwhile, the fairness of each user is defined based on the user's own provided feature. Bose and Hamilton [25] introduced compositional fairness for multiple sensitive attribute combinations for the knowledge graph embedding task. Li et al. [114] further defined personalized fairness based on causal notion for the collaborative filtering recommendation task. The system guarantees personalized counterfactual fairness for each user, i.e., the user's recommendation result is unchanged even if the personalized sensitive features that the user cares about were changed in a counterfactual world. For example, if the user does not want to be discriminated by the recommender system on his or her age, then the system can guarantee that the recommendations (such as news articles) are age-neutral or age-diverse. Wu et al. [191] further advanced from personalized fairness to personalized selective fairness in recommendation, which provides users with the flexibility to select the sensitive feature(s) that they care about after the recommendation model is trained, and thus users can freely re-select the sensitive feature(s) that they care about when they want.

5.3.8 ***Blackbox Fairness vs. Explainable Fairness.*** Most fairness in recommendation research focus on defining fairness and developing methods to improve the fairness. However, an even more fundamental problem is to understand why a model is unfair, i.e., what reasons lead to unfair model outputs. There have been research on explaining recommendation results [44, 76, 172, 196, 216, 217], explaining graph neural networks [122, 171, 195, 204, 205], explaining vision and language models [50, 82, 87, 110, 111], etc., but the research on explaining why a model is fair or unfair is still very limited. Understanding the "why" is not only helpful on technical perspectives but also on social perspectives. Technically, knowing the reasons of unfairness helps system designers to conduct data curation so as to remove the factors that lead to unfairness and also helps them to develop targeted models for disparity mitigation. Socially, knowing the reasons of unfairness helps policy makers to understand the social causes and implications of unfairness and develop prevention policies for future improvements. Explainable fairness is especially important for recommender systems because recommendation algorithms usually work with a huge amount (thousands or even millions) of both latent and explicit features in a collaborative learning way. In many other intelligent decision making systems such as loan approval and school admission, the total number of features is much smaller than recommender systems, and most importantly, it is usually straightforward for policy makers to know which feature(s) are sensitive that should be excluded from decision making so as to guarantee fairness, such as gender and race. However, recommender systems are different, on one hand, the huge amount of features makes it difficult to manually identify the sensitive features that lead to unfairness; on the other hand, most features are not immediately related to commonly known sensitive features such as race and gender but they still implicitly influence the model fairness; finally, even if the recommendation algorithm does not use explicit features at all and only uses user behaviors for model training, it may still result in unfair recommendations due to collaborative learning since some users' preferences and choices will influence other users' received recommendations. As a result, explainable fairness methods are highly demanded in recommender systems that can help to detect and explain why a model is unfair and how to improve. Ge et al. [74] aim to develop explainable fairness models for recommendation. The authors take item exposure fairness as an example and develop a counterfactual explainable fairness framework to explain which item features in the system significantly influence the model fairness. By reducing the influence of the detected sensitive features, the model is able to achieve better fairness-utility trade-off in recommendation.

5.4 Methods

Several methods have been proposed to mitigate unfairness in recommendation. The research on fair recommendation usually first define the fairness metrics they concern and then develop suitable techniques to promote the corresponding metrics. The relevant techniques are diverse and can be very different under different research and fairness definitions. However, we try to organize them into several categories in this section and show some typical and common-used methods in each category so as to help readers better understand how fairness is technically achieved.

5.4.1 ***Regularization and Constrained Optimization.*** So far, the techniques for promoting fairness in recommendation are mainly in the form of regularization and constrained optimization. The various fairness criteria can be formulated as regularizers or constraints to guide the process of model optimization [15, 30, 67, 112, 117, 202, 203, 219]. The objective can be maximizing the utility of recommendation under fairness constraints [202], or maximizing fairness requirements under utility bounds [219], or jointly optimizing both fairness and utility goals with a reasonable trade-off [72]. Existing works usually apply regularization or constraint optimization to in-processing methods [15, 202] and post-processing methods [77, 112]. The challenges for regularization and constrained optimization methods are that they are often non-convex in nature and are difficult to balance the conflicting constraints which may result in unstable training [35]. A very clear example is [202]. The authors propose four new unfairness metrics for preference prediction in collaborative filtering based recommendation, all measuring the discrepancy on prediction quality between the disadvantaged users and advantaged users. The disadvantaged and advantaged users are divided based on a binary group feature such as gender. Let's use $\mathrm{E}_g[r]_j$ and $\mathrm{E}_{\neg g}[r]_j$ to denote the average ratings for the j-th item from the disadvantaged and advantaged users, respectively. $\mathrm{E}_g[y]_j$ and $\mathrm{E}_{\neg g}[y]_j$ denote the average predicted score for the j-th item from disadvantaged users and advantaged users, respectively. We show the first fairness metric *Value Unfairness* here as an example, and it measures the inconsistency in signed estimation error across user groups as:

$$U_{\text{val}} = \frac{1}{n}\sum_{j=1}^{n}\left|\left(\mathrm{E}_g[y]_j - \mathrm{E}_g[r]_j\right) - \left(\mathrm{E}_{\neg g}[y]_j - \mathrm{E}_{\neg g}[r]_j\right)\right|$$

To promote fairness in recommendation, the learning objective of recommendation task is extended with a smoothed variation of the fairness metrics. For example, the outer absolute value is replaced with the squared difference. All the four fairness metrics have straightforward subgradients and can be optimized by various subgradient optimization techniques. The authors solve the problem for a local minimum as follows:

$$\min_{\theta} L_{Rec}(\theta) + U.$$

Other examples include [15], where the authors propose to measure fairness based on pairwise comparisons and offer a regularizer to encourage improving this metric during model training and thus improve fairness in the resulting rankings; In [117], the authors study unfairness issue in group recommendation, where items are recommended for a group of users whose preferences can be different from each other. Several fairness metrics are proposed for group recommendation scenario. The problem is formulated as a multiple objective optimization problem with the fairness metric as a regularizer and is solved from the perspective of Pareto Efficiency.

5.4.2 ***Adversary Learning.*** Another typical technique to promote fairness is to take advantage of adversary learning [10, 13, 17, 25, 59, 62, 64, 176, 197]. The basic idea of mitigating unfairness through adversary learning is to learn fair representations through a min-max game between the main task predictor and an adversarial classifier. The predictor aims to learn informative representations for the recommendation task, while the goal of the adversarial classifier is to minimize the predictor's ability to predict the sensitive features from the representations, and thus the information about sensitive features are removed from the representations to mitigate discrimination. For example, let's consider $\mathcal{L}_{Rec}$ to denote the loss of the recommendation task such as the pair-wise ranking loss [156] or the mean square error loss [104]. $\mathcal{L}_A$ denotes the loss of the adversarial classifier for sensitive feature S which can be a cross-entropy loss for implementation. The adversary learning loss can be as follows:

$$\mathcal{L} = \sum_{(u,v)\in\mathcal{D}} \Big(\mathcal{L}_{\mathrm{Rec}}(u, v, y_{uv}) - \lambda \cdot \mathcal{L}_A\,(u, S)\Big) \tag{13}$$

where the adversarial coefficient λ controls the trade-off between recommendation performance and fairness. An advantage of adversary approaches is that they often treat the prediction model as a black-box so as to offer the model-agnostic flexibility. However, the adversary approaches are usually difficult to train and lack stability [16, 69]. Beigi et al. [13] build a recommendation model with attribute protection to counter private-attribute inference attacks in recommendation. The authors solve the problem through adversarial learning with two main components: the private attribute inference attacker and the Bayesian personalized recommender. The goal of the attacker is to infer user private-attribute information while the recommender aims to learn users' interests under the regularization of the attacker. Wu et al. [186] propose a fairness-aware approach based on decomposed adversarial learning for news recommendation to mitigate the unfairness brought by the biases of user sensitive features. The proposed method decomposes user embedding into two parts: a bias-aware user embedding to capture the bias information on sensitive features and a bias-free user embedding for making fairness-aware news ranking. Adversary learning is used for removing the bias information of sensitive features from user embeddings.

5.4.3 ***Reinforcement Learning.*** Some researches believe that recommendation is not only a prediction problem but a sequential decision problem, and suggest to model the recommendation problem as a Markov Decision Process (MDP) and solve the problem through Reinforcement Learning (RL) [8]. Some research on fairness in recommendation also follow the trend and consider RL methods to promote fairness through a long-term and dynamic perspective [73, 75, 89, 93, 120, 180–182, 210]. In [73], the authors consider the dynamic item exposure fairness in recommender systems where the item popularity will change over time with recommendation actions and user feedback. The problem is formulated as a Constrained Markov Decision Process by dynamically constraining the item exposure fairness metric at each iteration. In [75], the authors study the problem of Pareto optimal fairness-utility trade-off in recommendation and propose a framework based on multi-objective reinforcement learning to learn a single parametric representation for optimal recommendation policies over the space of all possible preferences. In [120], the authors study the problem of interactive recommender systems. Considering that the user preferences and the system fairness status are constantly changing over time, the authors jointly represent the user preferences and the system's fairness status into the states of the MDP recommendation model through an RL framework to achieve long-term fairness dynamically. One challenge of leveraging RL to promote fairness is that there may be a risk of "user tampering" [65], which is a phenomenon that a recommender system may try to increase its long-term user engagement through manipulating the user's opinions, preferences and beliefs via its recommendations. The potential unfairness of users that different users are affected by different manipulation needs vigilance from researchers and system developers [210].

5.4.4 ***Causal Methods.*** Causal methods for promoting fairness have been widely studied in classification tasks [46, 71, 78, 102, 105, 136, 193] and have also recently been applied to the field of recommendation. The key objective of causal methods is to investigate the relationships underlying the data and model, including the causal effects between sensitive variables and decisions, as well as the dependency between sensitive and non-sensitive variables. For example, Wu et al. [192] study causal-based anti-discrimination method in ranking problem. The authors build a causal graph to identify and remove both direct and indirect discrimination in ranked data and reconstruct a fair ranking if discrimination is detected from the causal graph; Zheng et al. [218] leverage causal inference to solve popularity bias in recommendation. The authors assume that the user click behaviour depends on both their interest and the item popularity. To mitigate the effect of popularity bias, the authors propose a general framework to disentangle user interest and popularity bias through learning two types of embeddings: interest embedding and popularity embedding. The final recommendations are generated by only the interest embeddings so that the popularity bias is removed; Huang et al. [89] study how to achieve counterfactual fairness for users in online recommendation through incorporating causal inference into bandits. The authors adopt soft intervention to model the arm selection strategy and use the d-separation set identified from the causal graph to develop a fair UCB algorithm, which promotes fairness through choosing arms that satisfy the counterfactual fairness constraint. Another commonly-used causal-based methods is re-weighting the instances based on Inverse-Propensity-Scoring (IPS) techniques [158], which are usually adopted to solve the biases in recommendation such as popularity bias [200] and selection bias [162]. The IPS methods assume that the biases are caused by the fact that treatments are not being randomly assigned, and thus they use the inverse propensity of assigning the treatment as the sample weights to remove the bias. The IPS methods are usually convenient to implement, but the score estimator may not properly handle large shifts in observational probability [21] and the methods are usually specifically designed for a particular problem setting. Besides improving fairness, the causal methods can also provide model transparency in terms of how decisions are made [88]. A recent work [74] studies the problem of how to explain which feature(s) lead to item exposure unfairness in recommendation through counterfactual reasoning. Causal methods also have challenges in practice, for example, the causal-based notions are usually defined based on intervention and counterfactual which are non-observable quantities and cannot always be computed from observational data [125]; and causal methods usually require the prior knowledge of causal graphs or causal assumptions which may not always be accessible in practice [159].

5.4.5 ***Other Methods.*** Several other techniques are used to promote fairness of recommendations. Examples include 1) data augmentation methods such as [155], which proposes a strategy to improve the socially relevant properties such as individual or group fairness of a recommender system through adding antidote data. The proposed framework is developed from an existing pre-trained matrix-factorization-based recommender system, and provides the flexibility of not having to modify the original input data or recommendation algorithm. The authors consider to add the ratings of new users to the input which are chosen based on the corresponding measure so as to improve the fairness of recommender systems; 2) methods based on Variational Autoencoders (VAEs) such as [24], which proposes to incorporate randomness into the operation of VAEs so as to mitigate the position bias in multiple rounds of recommendation. The authors introduce four different noise distributions and find that adding noise to the process of sampling values from VAE's latent representation can provide long term fairness for recommendation with an acceptable trade-off between fairness and recommendation quality such as NDCG; 3) methods based on self-distillation [184], which leverages the model predictions on the original data as a teacher to regularize the predictions on the augmented data with randomly dropped user behaviors. The authors observe that the training process of big recommendation models can result in unfair recommendation performance for cold users, and show that the proposed self-distillation method can push the model to fairly capture the interest distributions of heavy and cold users. More techniques are used to improve fairness in other fields of machine learning [35, 175], but their application in the recommendation scenario is very limited.

5.5 Datasets for Fairness in Recommendation Research

In this section, we collect some datasets for fairness research in recommendation. The requirements for datasets in fairness research usually highly depend on the specific fairness definitions. For example, for fairness research on user side, researchers usually need to consider fairness definitions about users' sensitive features; while for fairness research on item side, datasets with item features such as item category may be preferred. Some works even do not consider the fairness on user/item sensitive features but instead on user activity or item popularity, which can be directly computed from the user interaction data, and thus they have no requirement for user/item features from the datasets. Compared with the datasets that contain user-item interactions and item features, recommendation datasets that contain user sensitive features are very limited and uneasy to find. Therefore, we introduce several datasets for recommendation where the user sensitive feature information is available. All datasets are publicly available under the referenced links.

5.5.1 ***MovieLens***[1]. This is a dataset for movie recommendation tasks with 1,000,209 anonymous user-movie ratings. It has approximately 3,900 movies and 6,040 users with each rating within 1 to 5. It also has a smaller version dataset with 100,000 ratings, 943 users and 1682 movies. There are three user sensitive features: gender, age and occupation.

5.5.2 ***Last.FM***[2]. This is a large dataset for music retrieval and recommendation tasks [131]. It has more than 2 billion user-music listening events, 120,322 users, 5,159,580 artists and 50,813,373 tracks. There are three user sensitive features: gender, age and country.

5.5.3 ***Last.FM- 360K***[3]. This is a smaller music recommendation dataset collected from the Last.fm website [37]. It has 17 million records with 36,000 users and 290,000 artists. There are three user sensitive features: gender, age and country.

5.5.4 ***RentTheRunWay***[4]. This dataset contains user-cloth renting interactions [133]. It has 192,544 interactions, 105,508 users and 5,850 items. User age is included as a sensitive feature.

5.5.5 ***Amazon Electronics***[5]. This dataset contains user online shopping behaviors on Amazon for e-commerce recommendation [174]. It has 1,292,954 shopping events, 1,157,633 users and 9,560 products. User gender information is included as a sensitive feature.

5.5.6 ***Alibaba***[6]. This datast is from the Alibaba e-commerce platform [147] which contains user-page interactions. It considers a recommendation scenario where a user is recommended with 50 items per page in a ranked order. There are 49 million users and 200 million items. The interaction

[1] https://grouplens.org/datasets/movielens/1m/

[2] http://www.cp.jku.at/datasets/LFM-2b/

[3] http://ocelma.net/MusicRecommendationDataset/lastfm-360K.html

[4] https://cseweb.ucsd.edu/~jmcauley/datasets.html#clothing_fit

[5] https://cseweb.ucsd.edu/~jmcauley/datasets.html#market_bias

[6] https://github.com/rec-agent/rec-rl

types include requests (refresh recommendation lists), clicks, adding to carts, adding to wishlists and purchases. Age and gender information is included as the sensitive features.

5.5.7 ***AliEC***[7]. This is a dataset of click rate prediction about display Ad from the Alibaba e-commerce platform. It contains about 100,000 users and 8.8 million click instances. Age and gender information is included as the sensitive features.

5.5.8 ***Insurance***[8]. This is a Kaggle dataset with the goal of recommending insurance products to a target user. It has 5,382 interactions, 1,231 users and 21 insurance policies. There are three user sensitive features: gender, marital status and occupation.

5.5.9 ***Post***[9]. This dataset is for post recommendation tasks. It has 71,800 user-post interactions, 500 users and 6,000 posts with user gender information as the sensitive feature.

5.5.10 ***Coat***[10]. This dataset is collected by Schnabel et al. [162]. They hire 290 Mechanical Turkers to select a set of items among 300 candidates and ask them to place ratings on selected items. Each user comes with 24 rated items in the training set and 16 items in the testing set. There are two sensitive features: gender and age.

5.5.11 ***Sushi***[11]. This dataset is collected by Kamishima [96] which includes responses of a questionnaire survey of preference in SUSHI. It can be used for rating prediction or ranking tasks. There are 5,000 users and 100 SUSHI candidates. This dataset contains a ranking data which asks users to rank the order of SUSHI based on their personal preferences and a scoring data which asked users to give 1 to 5 rating to the ordered food. There are two sensitive features: gender and age.

6 CHALLENGES AND OPPORTUNITIES

So far, researchers have realized the importance of improving fairness in recommender systems and have started the exploration. However, the research in this area are still very limited and many important problems have not even been studied. In this section, we talk about some open challenges and point out some future opportunities with the hope to inspire more research works in this area.

6.1 Lack of Consensus on Fairness Definitions

The most challenging fact in fairness research is that there is no unique consensus on fairness definition. The fairness considerations are different in different scenarios, the biases that lead to unfairness are diverse, and the fairness demands can be put forward from several different perspectives. The lack of consensus on fairness definition leads to a series of problems.

First, how to achieve multiple fairness requirements at the same time. Existing methods are usually designed for achieving one particular type of fairness requirement. However, this is not sufficient for solving the problem since people's demand for fairness is diverse and various biases often occur simultaneously [43]. As a result, it is imperative to explore whether it is possible to propose a unified model to handle multiple fairness demands. Though a single solution to all fairness problems is impossible since it has been theoretically proven that some fairness notions are inherently conflicting with each other and cannot be achieved at the same time [103], it is still worth it to explore some simple scenarios such as handling two or three different cases. Furthermore, it is interesting to explore the relationship between various biases and fairness notions

[7] https://tianchi.aliyun.com/dataset/dataDetail?dataId=56

[8] https://www.kaggle.com/mrmorj/insurance-recommendation

[9] https://www.kaggle.com/datasets/vatsalparsaniya/post-pecommendation

[10] https://www.cs.cornell.edu/~schnabts/mnar/

[11] https://www.kamishima.net/sushi/

and clarify whether we really need to address so many fairness definitions. It could be possible that as long as the system guarantees several important fairness notions then most other fairness notions will be naturally guaranteed or guaranteed to some extent, and thus we do not have to work on indefinitely many fairness notions. Second, when some fairness requirements cannot be achieved simultaneously due to their conflicts, one important problem is how to achieve a good and reasonable balance. Solving the problem may not only need technical innovations such as Pareto optimality but also careful social and ethical considerations such as transparency and trust. Third, if the system is designed for achieving only one fairness requirement, it is important to know how to choose the most suitable fairness consideration for the specific case. For example, which one is more important for improving user fairness in recommendation? Group fairness or individual fairness? Static fairness or dynamic fairness? To answer the questions, it is important to study how to evaluate different fairness notions for a specific scenario. This further leads to another challenge: since existing works are usually proposed under different perspectives, they often adopt different or even unique evaluation metrics to evaluate fairness. Therefore, it is sometimes difficult to compare the existing methods. To this end, some evaluation metrics or benchmark datasets that can help us to compare different fairness notions from a unified view are highly appreciated in the future.

6.2 Relationship between Fairness and other Trustworthy AI Perspectives

Fairness is a very important perspective for trustworthy, responsible and ethical AI research. However, except for fairness, there are many other perspectives to consider, such as explainability, controllability, robustness and privacy. A very important problem is to explore the relationship between fairness and other trustworthy AI perspectives and how can the different perspectives interact with each other to develop better trustworthy AI systems.

Take fairness and explainability as an example. As mentioned above, due to the inherent conflict between some fairness definitions, in some cases, the system may not be able to meet every user's fairness requirement due to the conflicting fairness demands, and in such cases, transparency and honesty from the system may be highly appreciated. For instance, the system may honestly explain to some users why they have to bear a bit of unfair treatment at the current moment in order to better serve other (usually the majority of) people or for certain long-term goals, and how the system will compensate such unfair treatment in the future. Besides, if we implement a fair system, we also need to explain to users why a decision is fair for the user so as to make sure users trust the intelligent system. This requires the research community to better understand the relationship between explainability and fairness and develop explainable fairness models in the future so that explainability and fairness can benefit each other. Other research topics include the relationship between fairness and robustness such as out-of-distribution generalization, the relationship between fairness and privacy such as the fairness in federated learning, and the relationship between fairness and controllability such as if and how users can actively control the fairness demands for themselves.

6.3 Causal Foundations for Fairness

As we introduced in section 5.3.4, it is important to consider causal-based fairness notions to address unfairness in recommendation. There are some works considering causality in classification [105, 109] and general recommendation [113, 198, 199] tasks, however, the exploration of causal fairness in recommendation problem is still very limited. Although recommendation task can be formulated as a classification problem in some cases, the causal fairness techniques in classification may not be directly migrated to the recommendation problem. For example, to achieve counterfactual fairness in classification, as stated in [105], the most straightforward way to guarantee the independence between the predicted results and the sensitive features is just to avoid using the sensitive features (and the features causally depend on the sensitive features) as input. However, this is not the case in recommendation scenarios: most of the collaborative learning-based recommendation algorithms such as collaborative filtering and collaborative reasoning are directly trained from user-item interaction information [63] which do not use any feature, no matter sensitive feature or non-sensitive feature. However, the model still generates unfair recommendations on user sensitive features even if the model does not directly use any feature as input [114]. The reason is that during collaborative learning, the model may capture the underlying relationships between user features and user behaviours that are inherently encoded into the training data, since user features may have causal impacts on user behaviors and preferences. As a result, we need to design methods to achieve counterfactually fair recommendations and this cannot be realized in trivial ways such as not using sensitive feature, because the collaborative learning model does not use sensitive feature from the beginning. Therefore, more explorations about causality considerations are needed to understand the underlying reasons of unfairness and solve the unfairness issue properly.

6.4 Understanding the Connection between Bias and Fairness

Bias and fairness are often mentioned concurrently or even obfuscated sometimes, but their relationship has not been clearly understood or discussed in the community. In general, various biases could be the main causes of the unfair results in recommendation. For example, the popularity bias in data can cause exposure unfairness on item side, while the gender bias in data can cause unfair treatment on user side. Besides the various biases in data or algorithm, there are also other reasons for unfairness. For example, researches have shown that some fairness requirements can not be satisfied at the same time [47, 103, 149], therefore, the violation of one type of fairness may be caused by ensuring another. Recently, researchers have explored various bias and debias methods in recommendation [43], however, existing debias researches usually focus on how to use debias to improve the recommendation accuracy and seldom consider to promote fairness. On the other hand, many works on fairness are not implemented through debiasing methods but directly adding fairness requirements over the model outcome, for example, by adding a fairness constraint on the optimization process, which bears the risk of losing accuracy. Therefore, there is a gap between the research on debias and fairness though the two problems have deep connections both theoretically and practically. As a result, the relationship between bias and unfairness should be carefully considered and the connection between debiasing and fairness promotion methods should be better established, which may lead to better understandings on the causes of unfairness as well as better methods for improving both fairness and accuracy.

6.5 Considerations in Real-World Deployment

Although many fairness research have been conducted in academia and industry, deploying fairness-aware models to benefit users in real-world system requires a lot more practical considerations. The deployment of fairness-aware models in industry can be broadly classified into two types: user-oriented deployment and developer-oriented deployment. User-oriented deployment focus on delivering fairness-aware results to real users of the system and thus directly influence the service of users, while developer-oriented deployment do not immediately deliver such results to users but instead focus on developing tools within the production environment to help developers and policy makers better understand the system unfairness. It is worth noting that the "users" in user-oriented deployment not only include the average users of the system such as the consumers in e-commerce or the viewers in short video social networks, other stakeholders in the system such as the various sellers in e-commerce and the video creators in social networks are also users of the platform, who also have their fairness demands such as fair exposure opportunity of their items.

An example of user-oriented deployment is LinkedIn [77], which explored the first large-scale deployed framework for ensuring fairness in the hiring domain. Researchers applied the fairness framework to LinkedIn Talent Search so as to achieve fairness criteria such as equality of opportunity and demographic parity in candidate ranking. Researchers also presented the online A/B testing results, which showed that their approach resulted in tremendous improvement in fairness metrics without affecting the business metrics. An example of developer-oriented deployment is Amazon SageMaker Clarify [84], which developed explainability toolkits for Amazon SageMaker and was deployed in the Amazon AWS cloud computing clusters. Based on the toolkits, various developers over the world who use the cloud cluster can easily detect and monitor if there is any bias in their data or model and if the results produced by their model are fair. This can help the developers to better understand the consequences of their models and thus help them to better refine the models. Due to the multi-stakeholder nature of fairness in real-world systems [1], in the future, it is important to understand the difference and relationship between the various fairness demands from various stakeholders in real-world systems, and how the knowledge from developer-oriented deployment can be transferred to user-oriented deployment so as to directly benefit the users.

6.6 Understanding the Relationship between Fairness and Utility

As mentioned above, fairness-aware methods may be deployed in user-oriented or developer-oriented ways in practice. Developer-oriented deployment usually meet fewer obstacles in real-world systems because it does not immediately influence the users and business metrics. However, more in-depth considerations or trade-offs usually need to be considered for user-oriented deployment in practical systems. On one hand, real-word recommender systems are usually very complex with multiple modules and massive training data, and thus it can be difficult to figure out all of the fairness problems in such a huge system. On the other hand, many industry companies usually put business metrics such as purchase rates first compared to the fair treatment of users, and thus the incentive to promote fairness is less than that of chasing profits in many practical systems, especially if there is an inevitable trade-off between fairness and profit metrics. Of course, one approach to promoting fairness in practice is through various legal requirements that are already implemented in many countries, such as the General Data Protection Regulation (GDPR) in EU[12], California Consumer Privacy Act (CCPA) in the US[13], and the Internet Information Service and Algorithmic Recommendation Regulation (IISARR) in China[14], which enforce fair treatment of users in real-world systems. However, on the other hand, it is also very important for the research community to further explore the relationship between fairness and utility so as to improve the incentive of industry practitioners to promote fairness.

This requires further explorations on three folds: 1) Explore the cases where fairness and utility do not conflict with each other but actually promote each other. Actually, there have been research on user-oriented group-fairness in recommendation [112, 154] which show that by carefully defining the groups and fairness metrics, it is possible for recommendation models to improve both fairness and utility (such as recommendation accuracy) at the same time. In the future, it is promising to explore if co-improving fairness and utility is possible under other fairness and utility definitions; 2) If there is really an inevitable conflict between certain fairness and utility metrics, we can develop methods to guarantee fairness with as few utility losses as possible or guarantee utility with as few fairness losses as possible, explain such losses to users in appropriate ways so that they can accept it, and amortize the losses across users or time so that the system does not always hurt some (types of)

[12] https://gdpr.eu/

[13] https://oag.ca.gov/privacy/ccpa

[14] http://www.cac.gov.cn/2022-01/04/c_1642894606364259.htm

users. This may require further innovations on multi-objective optimization, Pareto optimization, explainable fairness, dynamic fairness and creative user interfaces in the future; 3) Same as any other trustworthy AI research, understanding the benefits of fairness in recommendation should be considered in the dynamic and long-term context. Some research show that certain fairness and utility metrics may have to trade-off with each other in short-term settings [75, 179], however, due to the limitations on datasets and evaluation methods, existing research on fairness-utility relationship are mostly conducted under static environments using fixed datasets rather than in dynamic deployed systems. As a result, it is not surprising to observe that increasing fairness may hurt recommendation accuracy because recommending more long-tail items to increase exposure fairness will naturally lead to fewer recommendation hits in static datasets where popularity bias of user clicks are already recorded in the testing set. This does not necessarily mean that increasing fairness definitely hurts utility in deployed systems, because in the long-term, if users feel they are fairly treated by the system, it will help to increase users' retention, interest, trust and engagement in the system, and will thus help to create a sustainable eco-system in the platform, which eventually leads to increased utility and profits for the platform in the long-run. As a result, it is important to explore the relationship between utility and fairness in the long-run in real-world deployed systems, and this requires close collaboration between academia and industry in the future.

6.7 Developing Better Metrics, Benchmarks and Simulation Environments

Though we have introduced several benchmark datasets to study fairness in this survey, the available dataset to support fairness research is still very limited. This is because the research on fairness usually requires datasets that have sufficient features such as user personal information. Besides, there exist many different fairness concerns and each requires certain dataset to study, which means that a diverse portfolio of datasets are needed for fairness research. As a result, extensive efforts are needed in the future to create various datasets to support fairness study. More importantly, this may require innovative solutions to meet the demands of both researchers and data providers. For example, many fairness research rely on users' personal information, however, most of the personal information are users' sensitive features such as gender, race, age, location and income that are closely related to users' privacy. As a result, researchers should take extreme care when creating and publishing datasets for fairness research such as proper anonymization of users and even developing advanced methods for both privacy protection and fairness promotion, such as through federated learning or differential privacy.

On the other hand, many users in real-world systems may choose to not reveal their personal information in registration or even provide wrong personal information. The reason for users to do so is exactly because they are afraid of being unfairly treated by the system or breach of their privacy. However, without such information, it would be difficult to improve fairness for them even if the system want to do so, which results in a dilemma due to the crisis of confidence. As a result, it is important to create a good incentive and build trust with users so that they are willing to provide the relevant information to support fairness promotion. Besides, the community can explore reliable data augmentation and data curation methods to improve the quality of incomplete datasets, and develop fairness-aware methods that do not rely on users' sensitive features when they are not available.

Fairness evaluation is also an important problem that needs further exploration. As mentioned in previous sections, the fairness definitions are very diverse. Currently, each type of fairness has its unique evaluation methods, which makes it difficult to compare different fairness methods or develop unified models that can improve fairness from multiple perspectives. As a result, it would be promising if a systematic evaluation protocol can be developed which enables us to evaluate diffident fairness requirements on the same scale. Furthermore, we may even develop methods to evaluate various utility and trustworthy considerations in the same framework, such as fairness, accuracy, explainability, robustness and privacy. Finally, to support the evaluation of dynamic fairness and to explore the long-term effect of fairness, a simulation environments will be highly appreciated to the community. Researchers have attempted to build simulation platforms for recommendation accuracy research such as RecSim [90], and similar platforms for fairness research will be very useful not only to the recommender system community but also to the broader fairness in AI communities.

7 CONCLUSIONS

In this survey, we aim to introduce the foundations, definitions, methods and outlooks for fairness in recommendation research. We begin the survey with a brief introduction of fairness in machine learning to provide beginners with a general view and basic background knowledge of fairness research, including the causes, methods, as well as different considerations of fairness. To help better understand the fairness concepts in recommendation, we also introduce the basic fairness definitions and methods in classification and ranking tasks since they are closely related to fairness in recommender systems. For the main body of the survey, we provide a taxonomy to classify the existing fairness definitions in recommender systems to help reader build a systematic understanding of the area. We also introduce the various technical and evaluation methods as well datasets for fairness in recommendation research. Finally, we discuss the challenges and opportunities of fairness research in recommendation with the hope of both inspiring future innovations and promoting fairness deployment in real-world systems.

REFERENCES

[1] Himan Abdollahpouri and Robin Burke. 2019. Multi-stakeholder recommendation and its connection to multi-sided fairness. In *Proceedings of the RMSE workshop held in conjunction with the 13th ACM Conference on Recommender Systems*.

[2] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *RecSys*. 42–46.

[3] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing popularity bias in recommender systems with personalized re-ranking. *arXiv preprint arXiv:1901.07555* (2019).

[4] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The unfairness of popularity bias in recommendation. *arXiv preprint arXiv:1907.13286* (2019).

[5] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2020. The connection between popularity bias, calibration, and fairness in recommendation. In *Fourteenth ACM conference on recommender systems*. 726–731.

[6] Gediminas Adomavicius and YoungOk Kwon. 2011. Improving aggregate recommendation diversity using ranking-based techniques. *TKDE* (2011), 896–911.

[7] Gediminas Adomavicius and Alexander Tuzhilin. 2011. Context-aware recommender systems. In *Recommender systems handbook*. Springer, 217–253.

[8] M Mehdi Afsar, Trafford Crump, and Behrouz Far. 2021. Reinforcement learning based recommender systems: A survey. *arXiv preprint arXiv:2101.06286* (2021).

[9] Nil-Jana Akpinar, Cyrus DiCiccio, Preetam Nandy, and Kinjal Basu. 2022. Long-term Dynamics of Fairness Intervention in Connection Recommender Systems. In *Proceedings of the Conference on Artificial Intelligence, Ethics, and Society (AIES 2022)*.

[10] Mario Arduini, Lorenzo Noci, Federico Pirovano, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. 2020. Adversarial Learning for Debiasing Knowledge Graph Embeddings. *arXiv preprint arXiv:2006.16309* (2020).

[11] Abolfazl Asudeh, HV Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*. 1259–1276.

[12] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[13] Ghazaleh Beigi, Ahmadreza Mosallanezhad, Ruocheng Guo, Hamidreza Alvari, Alexander Nou, and Huan Liu. 2020. Privacy-aware recommendation with private-attribute protection using adversarial learning. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 34–42.

[14] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[15] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *SIGKDD*. 2212–2220.

[16] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[17] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

[18] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*. 405–414.

[19] Arpita Biswas, Gourab K Patro, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. 2021. Toward Fair Recommendation in Two-sided Platforms. *ACM Transactions on the Web (TWEB)* 16, 2 (2021), 1–34.

[20] 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).

[21] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *Proceedings of the 12th ACM conference on recommender systems*. 104–112.

[22] Ludovico Boratto, Gianni Fenu, and Mirko Marras. 2021. Interplay between upsampling and regularization for provider fairness in recommender systems. *User Modeling and User-Adapted Interaction* 31, 3 (2021), 421–455.

[23] Ludovico Boratto, Gianni Fenu, Mirko Marras, and Giacomo Medda. 2022. Consumer Fairness in Recommender Systems: Contextualizing Definitions and Mitigations. In *Proceedings of the 44th European Conference on Information Retrieval*.

[24] Rodrigo Borges and Kostas Stefanidis. 2019. Enhancing long term fairness in recommendations with variational autoencoders. In *Proceedings of the 11th international conference on management of digital ecosystems*. 95–102.

[25] Avishek Bose and William Hamilton. 2019. Compositional fairness constraints for graph embeddings. In *International Conference on Machine Learning*. PMLR, 715–724.

[26] Amanda Bower, Hamid Eftekhari, Mikhail Yurochkin, and Yuekai Sun. 2021. Individually Fair Ranking. *arXiv preprint arXiv:2103.11023* (2021).

[27] Eric Budish. 2011. The combinatorial assignment problem: Approximate competitive equilibrium from equal incomes. *Journal of Political Economy* 119, 6 (2011), 1061–1103.

[28] Robin Burke. 2002. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction* 12, 4 (2002), 331–370.

[29] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[30] Robin Burke, Nasim Sonboli, Masoud Mansoury, and Aldo Ordoñez-Gauger. 2017. Balanced neighborhoods for fairness-aware collaborative recommendation. (2017).

[31] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *Conference on fairness, accountability and transparency*. PMLR, 202–214.

[32] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*. IEEE, 13–18.

[33] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. 2017. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems* 30 (2017).

[34] Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports* 12, 1 (2022), 1–21.

[35] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).

[36] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2017. Ranking with fairness constraints. *arXiv preprint arXiv:1704.06840* (2017).

[37] Òscar Celma Herrada et al. 2009. *Music recommendation and discovery in the long tail*. Universitat Pompeu Fabra.

[38] Abhijnan Chakraborty, Gourab K Patro, Niloy Ganguly, Krishna P Gummadi, and Patrick Loiseau. 2019. Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 129–138.

[39] Harshal A Chaudhari, Sangdi Lin, and Ondrej Linda. 2020. A General Framework for Fairness in Multistakeholder Recommendations. *arXiv preprint arXiv:2009.02423* (2020).

[40] Hanxiong Chen, Yunqi Li, Shaoyun Shi, Shuchang Liu, He Zhu, and Yongfeng Zhang. 2022. Graph collaborative reasoning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 75–84.

[41] Haoyu Chen, Wenbin Lu, Rui Song, and Pulak Ghosh. 2020. Counterfactual Fairness through Data Preprocessing. (2020).

[42] Hanxiong Chen, Shaoyun Shi, Yunqi Li, and Yongfeng Zhang. 2021. Neural Collaborative Reasoning. In *WWW*.

[43] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and debias in recommender system: A survey and future directions. *arXiv preprint arXiv:2010.03240* (2020).

[44] Xu Chen, Yongfeng Zhang, and Ji-Rong Wen. 2022. Measuring" Why" in Recommender Systems: a Comprehensive Survey on the Evaluation of Explainable Recommendation. *arXiv preprint arXiv:2202.06466* (2022).

[45] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.

[46] Silvia Chiappa and William S Isaac. 2018. A causal Bayesian networks viewpoint on fairness. In *IFIP International Summer School on Privacy and Identity Management*. Springer, 3–20.

[47] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[48] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[49] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[50] Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. 2018. Automatic generation of natural language explanations. In *Proceedings of the 23rd international conference on intelligent user interfaces companion*. 1–2.

[51] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. 87–94.

[52] Elliot Creager, David Madras, Toniann Pitassi, and Richard Zemel. 2020. Causal modeling for fairness in dynamical systems. In *International Conference on Machine Learning*. PMLR, 2185–2195.

[53] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data* 5, 2 (2017), 120–134.

[54] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534.

[55] Pietro G Di Stefano, James M Hickey, and Vlasios Vasileiou. 2020. Counterfactual fairness: removing direct effects through regularization. *arXiv preprint arXiv:2002.10774* (2020).

[56] 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.

[57] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. 2021. Online certification of preference-based fairness for personalized recommender systems. *arXiv preprint arXiv:2104.14527* (2021).

[58] Qiang Dong, Shuang-Shuang Xie, and Wen-Jun Li. 2021. User-item matching for recommendation fairness. *IEEE Access* 9 (2021), 130389–130398.

[59] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems* (2020).

[60] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[61] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*. PMLR, 119–133.

[62] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).

[63] Michael D Ekstrand, John T Riedl, and Joseph A Konstan. 2011. *Collaborative filtering recommender systems*. Now Publishers Inc.

[64] Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. *arXiv preprint arXiv:1808.06640* (2018).

[65] Charles Evans and Atoosa Kasirzadeh. 2021. User Tampering in Reinforcement Learning Recommender Systems. *arXiv preprint arXiv:2109.04083* (2021).

[66] Francesco Fabbri, Maria Luisa Croci, Francesco Bonchi, and Carlos Castillo. 2022. Exposure Inequality in People Recommender Systems: The Long-Term Effects. *ICWSM* (2022).

[67] Golnoosh Farnadi, Pigi Kouki, Spencer K Thompson, Sriram Srinivasan, and Lise Getoor. 2018. A fairness-aware hybrid recommender system. *arXiv preprint arXiv:1809.09030* (2018).

[68] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[69] Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang. 2019. Learning fair representations via an adversarial framework. *arXiv preprint arXiv:1904.13341* (2019).

[70] 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.

[71] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint meeting on foundations of software engineering*. 498–510.

[72] Ruoyuan Gao and Chirag Shah. 2019. How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*. 229–236.

[73] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. 2021. Towards Long-term Fairness in Recommendation. *WSDM* (2021).

[74] 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. *SIGIR* (2022).

[75] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. 2022. Toward Pareto Efficient Fairness-Utility Trade-off inRecommendation through Reinforcement Learning. *WSDM* (2022).

[76] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). *arXiv preprint arXiv:2203.13366* (2022).

[77] 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.

[78] Bruce Glymour and Jonathan Herington. 2019. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In *Proceedings of the conference on fairness, accountability, and transparency*. 269–278.

[79] Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. 2016. Satisfying real-world goals with dataset constraints. *Advances in Neural Information Processing Systems* 29 (2016).

[80] David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. 1992. Using collaborative filtering to weave an information tapestry. *Commun. ACM* (1992).

[81] Elizabeth Gómez, Carlos Shui Zhang, Ludovico Boratto, Maria Salamó, and Mirko Marras. 2021. The Winner Takes it All: Geographic Imbalance and Provider (Un) fairness in Educational Recommender Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1808–1812.

[82] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual visual explanations. In *International Conference on Machine Learning*. PMLR, 2376–2384.

[83] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. 2020. Adversarial learning for counterfactual fairness. *arXiv preprint arXiv:2008.13122* (2020).

[84] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, et al. 2021. Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2974–2983.

[85] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[86] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. 173–182.

[87] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 264–279.

[88] Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. 2014. A peek into the black box: exploring classifiers by randomization. *Data mining and knowledge discovery* 28, 5 (2014), 1503–1529.

[89] Wen Huang, Lu Zhang, and Xintao Wu. 2021. Achieving Counterfactual Fairness for Causal Bandit. *arXiv preprint arXiv:2109.10458* (2021).

[90] Eugene Ie, Chih-wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. 2019. Recsim: A configurable simulation platform for recommender systems. *arXiv preprint arXiv:1909.04847* (2019).

[91] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[92] Rashidul Islam, Kamrun Naher Keya, Ziqian Zeng, Shimei Pan, and James Foulds. 2021. Debiasing career recommendations with neural fair collaborative filtering. In *Proceedings of the Web Conference 2021*. 3779–3790.

[93] Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2017. Fairness in reinforcement learning. In *International conference on machine learning*. PMLR, 1617–1626.

[94] Christopher Jung, Sampath Kannan, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. 2020. Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*. 677–678.

[95] Faisal Kamiran and Toon Calders. 2010. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*. Citeseer, 1–6.

[96] Toshihiro Kamishima. 2003. Nantonac collaborative filtering: recommendation based on order responses. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 583–588.

[97] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 35–50.

[98] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2014. Correcting Popularity Bias by Enhancing Recommendation Neutrality. In *RecSys*.

[99] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2018. Recommendation independence. In *Conference on fairness, accountability and transparency*. PMLR, 187–201.

[100] Mesut Kaya, Derek Bridge, and Nava Tintarev. 2020. Ensuring fairness in group recommendations by rank-sensitive balancing of relevance. In *Fourteenth ACM Conference on Recommender Systems*. 101–110.

[101] A. Khademi, S. Lee, D. Foley, and V. Honavar. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *WWW*.

[102] Niki Kilbertus, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. 2020. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 277–287.

[103] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Vol. 67. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 43.

[104] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.

[105] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[106] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. 2019. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 ieee 35th international conference on data engineering (icde)*. IEEE, 1334–1345.

[107] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User Fairness in Recommender Systems. In *Companion Proceedings of the The Web Conf.* 101–102.

[108] 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*. 297–307.

[109] Jiuyong Li, Weijia Zhang, Lin Liu, Kui Yu, Thuc Duy Le, and Jixue Liu. 2021. A general framework for causal classification. *International Journal of Data Science and Analytics* 11, 2 (2021), 127–139.

[110] Lei Li, Yongfeng Zhang, and Li Chen. 2020. Generate neural template explanations for recommendation. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 755–764.

[111] Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized Transformer for Explainable Recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 4947–4957.

[112] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. *WWW* (2021).

[113] Yunqi Li, Hanxiong Chen, Juntao Tan, and Yongfeng Zhang. 2022. Causal Factorization Machine for Robust Recommendation. In *Proceedings of the 2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

[114] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards Personalized Fairness based on Causal Notion. *SIGIR* (2021).

[115] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. 2021. CIKM 2021 Tutorial on Fairness of Machine Learning in Recommender Systems. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4857–4860.

[116] Yunqi Li, Yingqiang Ge, and Yongfeng Zhang. 2021. Tutorial on Fairness of Machine Learning in Recommender Systems. SIGIR.

[117] Xiao Lin, Min Zhang, Yongfeng Zhang, Zhaoquan Gu, Yiqun Liu, and Shaoping Ma. 2017. Fairness-aware group recommendation with pareto-efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 107–115.

[118] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: short-term attention/memory priority model for session-based recommendation. In *SIGKDD*. ACM, 1831–1839.

[119] Weiwen Liu and Robin Burke. 2018. Personalizing fairness-aware re-ranking. *arXiv preprint arXiv:1809.02921* (2018).

[120] Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. 2021. Balancing Accuracy and Fairness for Interactive Recommendation with Reinforcement Learning. *arXiv preprint arXiv:2106.13386* (2021).

[121] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2011. Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*. Springer, 73–105.

[122] Ana Lucic, Maartje A Ter Hoeve, Gabriele Tolomei, Maarten De Rijke, and Fabrizio Silvestri. 2022. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 4499–4511.

[123] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 502–510.

[124] Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. 2022. Learning Fair Node Representations with Graph Counterfactual Fairness. *arXiv preprint arXiv:2201.03662* (2022).

[125] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2020. Survey on Causal-based Machine Learning Fairness Notions. *arXiv preprint arXiv:2010.09553* (2020).

[126] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 2145–2148.

[127] Mirko Marras, Ludovico Boratto, Guilherme Ramos, and Gianni Fenu. 2021. Equality of learning opportunity via individual fairness in personalized recommendations. *International Journal of Artificial Intelligence in Education* (2021), 1–49.

[128] Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. 2021. Attributing Fair Decisions with Attention Interventions. *arXiv preprint arXiv:2109.03952* (2021).

[129] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[130] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*. 2243–2251.

[131] Alessandro B Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management* 58, 5 (2021), 102666.

[132] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 107–118.

[133] Rishabh Misra, Mengting Wan, and Julian McAuley. 2018. Decomposing fit semantics for product size recommendation in metric spaces. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 422–426.

[134] Aadi Swadipto Mondal, Rakesh Bal, Sayan Sinha, and Gourab K Patro. 2021. Two-sided fairness in non-personalised recommendations. *AAAI 2021 Student Abstract and Poster Program* (2021).

[135] 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.

[136] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[137] Mohammadmehdi Naghiaei, Hossein A Rahmani, and Yashar Deldjoo. 2022. CPFair: Personalized Consumer and Producer Fairness Re-ranking for Recommender Systems. *arXiv preprint arXiv:2204.08085* (2022).

[138] Preetam Nandy, Cyrus Diciccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. 2020. Achieving fairness via post-processing in web-scale recommender systems. *arXiv preprint arXiv:2006.11350* (2020).

[139] Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Serena Wang. 2020. Pairwise fairness for ranking and regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5248–5255.

[140] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[141] Tiago Palma Pagano, Rafael Bessa Loureiro, Maira Matos Araujo, Fernanda Vitoria Nascimento Lisboa, Rodrigo Matos Peixoto, Guilherme Aragao de Sousa Guimaraes, Lucas Lisboa dos Santos, Gustavo Oliveira Ramos Cruz, Ewerton Lopes Silva de Oliveira, Marco Cruz, et al. 2022. Bias and unfairness in machine learning models: a systematic literature review. *arXiv preprint arXiv:2202.08176* (2022).

[142] 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 *WWW*. 1194–1204.

[143] Gourab K Patro, Abhijnan Chakraborty, Niloy Ganguly, and Krishna Gummadi. 2020. Incremental fairness in two-sided market platforms: On smoothly updating recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 181–188.

[144] Gourab K Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair ranking: a critical review, challenges, and future directions. *arXiv preprint arXiv:2201.12662* (2022).

[145] Judea Pearl. 2009. *Causality*. Cambridge university press.

[146] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 560–568.

[147] Changhua Pei, Xinru Yang, Qing Cui, Xiao Lin, Fei Sun, Peng Jiang, Wenwu Ou, and Yongfeng Zhang. 2019. Value-aware recommendation based on reinforced profit maximization in e-commerce systems. *WWW* (2019).

[148] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2021. Fairness in rankings and recommendations: An overview. *The VLDB Journal* (2021), 1–28.

[149] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. *Advances in neural information processing systems* 30 (2017).

[150] Agoritsa Polyzou, Maria Kalantzi, and George Karypis. 2021. FaiREO: User Group Fairness for Equality of Opportunity in Course Recommendation. *arXiv preprint arXiv:2109.05931* (2021).

[151] Tao Qi, Fangzhao Wu, Chuhan Wu, Peijie Sun, Le Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2022. ProFairRec: Provider Fairness-aware News Recommendation. *arXiv preprint arXiv:2204.04724* (2022).

[152] Novi Quadrianto and Viktoriia Sharmanska. 2017. Recycling privileged learning and distribution matching for fairness. *Advances in Neural Information Processing Systems* 30 (2017).

[153] Tai Le Quy, Arjun Roy, Vasileios Iosifidis, and Eirini Ntoutsi. 2021. A survey on datasets for fairness-aware machine learning. *arXiv preprint arXiv:2110.00530* (2021).

[154] Hossein A Rahmani, Mohammadmehdi Naghiaei, Mahdi Dehghan, and Mohammad Aliannejadi. 2022. Experiments on Generalizability of User-Oriented Fairness in Recommender Systems. *SIGIR* (2022).

[155] 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.

[156] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[157] Yehezkel S Resheff, Yanai Elazar, Moni Shahar, and Oren Sar Shalom. 2018. Privacy and fairness in recommender systems via adversarial training of user representations. *arXiv preprint arXiv:1807.03521* (2018).

[158] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[159] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*. 793–810.

[160] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.

[161] Ryoma Sato. 2022. Enumerating Fair Packages for Group Recommendations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 870–878.

[162] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*. PMLR, 1670–1679.

[163] Qijie Shen, Wanjie Tao, Jing Zhang, Hong Wen, Zulong Chen, and Quan Lu. 2021. SAR-Net: A Scenario-Aware Ranking Network for Personalized Fair Recommendation in Hundreds of Travel Scenarios. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4094–4103.

[164] Shaoyun Shi, Hanxiong Chen, Weizhi Ma, Jiaxin Mao, Min Zhang, and Yongfeng Zhang. 2020. Neural Logic Reasoning. In *CIKM*. 1365–1374.

[165] Yuxin Shi, Han Yu, and Cyril Leung. 2021. A Survey of Fairness-Aware Federated Learning. *arXiv preprint arXiv:2111.01872* (2021).

[166] Ilya Shpitser and Judea Pearl. 2012. Identification of conditional interventional distributions. *arXiv preprint arXiv:1206.6876* (2012).

[167] 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.

[168] Ashudeep Singh and Thorsten Joachims. 2019. Policy learning for fairness in ranking. *Advances in Neural Information Processing Systems* 32 (2019).

[169] Ryosuke Sonoda. 2021. A Pre-processing Method for Fairness in Ranking. *arXiv preprint arXiv:2110.15503* (2021).

[170] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* 2 (2019).

[171] 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.

[172] 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.

[173] Sotiris Tsioutsiouliklis, Evaggelia Pitoura, Konstantinos Semertzidis, and Panayiotis Tsaparas. 2022. Link Recommendations for PageRank Fairness. In *Proceedings of the ACM Web Conference 2022*. 3541–3551.

[174] Mengting Wan, Jianmo Ni, Rishabh Misra, and Julian McAuley. 2020. Addressing marketing bias in product recommendations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 618–626.

[175] Mingyang Wan, Daochen Zha, Ninghao Liu, and Na Zou. 2021. Modeling Techniques for Machine Learning Fairness: A Survey. *arXiv preprint arXiv:2111.03015* (2021).

[176] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5310–5319.

[177] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. 2021. Deconfounded recommendation for alleviating bias amplification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1717–1725.

[178] Xuezhi Wang, Nithum Thain, Anu Sinha, Flavien Prost, Ed H Chi, Jilin Chen, and Alex Beutel. 2021. Practical Compositional Fairness: Understanding Fairness in Multi-Component Recommender Systems. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 436–444.

[179] Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. 2021. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 1748–1757.

[180] Min Wen, Osbert Bastani, and Ufuk Topcu. 2019. Fairness with dynamics. (2019).

[181] Min Wen, Osbert Bastani, and Ufuk Topcu. 2021. Algorithms for fairness in sequential decision making. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1144–1152.

[182] Joshua Williams and J Zico Kolter. 2019. Dynamic modeling and equilibria in fair decision making. *arXiv preprint arXiv:1911.06837* (2019).

[183] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. 2017. Learning non-discriminatory predictors. In *Conference on Learning Theory*. PMLR, 1920–1953.

[184] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. Are Big Recommendation Models Fair to Cold Users? *arXiv preprint arXiv:2202.13607* (2022).

[185] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022. FairRank: Fairness-aware Single-tower Ranking Framework for News Recommendation. *arXiv preprint arXiv:2204.00541* (2022).

[186] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairrec: fairness-aware news recommendation with decomposed adversarial learning. AAAI.

[187] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. 2021. Multi-FR: A Multi-Objective Optimization Method for Achieving Two-sided Fairness in E-commerce Recommendation. *arXiv preprint arXiv:2105.02951* (2021).

[188] Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. 2022. Joint Multisided Exposure Fairness for Recommendation. *arXiv preprint arXiv:2205.00048* (2022).

[189] Le Wu, Lei Chen, Pengyang Shao, Richang Hong, Xiting Wang, and Meng Wang. 2021. Learning fair representations for recommendation: A graph-based perspective. In *Proceedings of the Web Conference 2021*. 2198–2208.

[190] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. 2021. Tfrom: A two-sided fairness-aware recommendation model for both customers and providers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1013–1022.

[191] Yiqing Wu, Ruobing Xie, Yongchun Zhu, Fuzhen Zhuang, Xiang Ao, Xu Zhang, Leyu Lin, and Qing He. 2022. Selective Fairness in Recommendation via Prompts. *SIGIR* (2022).

[192] Yongkai Wu, Lu Zhang, and Xintao Wu. 2018. On discrimination discovery and removal in ranked data using causal graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2536–2544.

[193] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

[194] Y. Wu, L. Zhang, X. Wu, and H. Tong. 2019. Pc-fairness: A unified framework for measuring causality-based fairness. *arXiv preprint arXiv:1910.12586* (2019).

[195] Yikun Xian, Zuohui Fu, Shan Muthukrishnan, Gerard De Melo, and Yongfeng Zhang. 2019. Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval.* 285–294.

[196] Yikun Xian, Tong Zhao, Jin Li, Jim Chan, Andrey Kan, Jun Ma, Xin Luna Dong, Christos Faloutsos, George Karypis, Shan Muthukrishnan, et al. 2021. Ex3: Explainable attribute-aware item-set recommendations. In *Fifteenth ACM Conference on Recommender Systems.* 484–494.

[197] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. *arXiv preprint arXiv:1705.11122* (2017).

[198] Shuyuan Xu, Yingqiang Ge, Yunqi Li, Zuohui Fu, Xu Chen, and Yongfeng Zhang. 2021. Causal Collaborative Filtering. *arXiv preprint arXiv:2102.01868* (2021).

[199] Shuyuan Xu, Juntao Tan, Shelby Heinecke, Jia Li, and Yongfeng Zhang. 2021. Deconfounded Causal Collaborative Filtering. *arXiv preprint arXiv:2110.07122* (2021).

[200] Longqi Yang, Yin Cui, Yuan Xuan, Chenyang Wang, Serge Belongie, and Deborah Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM conference on recommender systems.* 279–287.

[201] Tao Yang, Zhichao Xu, and Qingyao Ai. 2022. Effective Exposure Amortizing for Fair Top-k Recommendation. *arXiv preprint arXiv:2204.03046* (2022).

[202] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems.* 2921–2930.

[203] Sirui Yao and Bert Huang. 2017. New fairness metrics for recommendation that embrace differences. *arXiv preprint arXiv:1706.09838* (2017).

[204] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. 2019. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems* 32 (2019).

[205] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. 2020. Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.* 430–438.

[206] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.

[207] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management.* 1569–1578.

[208] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020.* 2849–2855.

[209] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000* (2021).

[210] Dell Zhang and Jun Wang. 2021. Recommendation Fairness: From Static to Dynamic. *arXiv preprint arXiv:2109.03150* (2021).

[211] Junzhe Zhang and Elias Bareinboim. 2018. Equality of opportunity in classification: A causal approach. In *NeurIPS.* 3675–3685.

[212] Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *AAAI*, Vol. 32.

[213] Xueru Zhang and Mingyan Liu. 2021. Fairness in learning-based sequential decision algorithms: A survey. In *Handbook of Reinforcement Learning and Control.* Springer, 525–555.

[214] Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. 2020. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems* 33 (2020), 18457–18469.

[215] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W Bruce Croft. 2017. Joint representation learning for top-n recommendation with heterogeneous information sources. In *CIKM.*

[216] Yongfeng Zhang and Xu Chen. 2020. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval* 14, 1 (2020), 1–101.

[217] 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.

[218] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021.* 2980–2991.

[219] Ziwei Zhu, Xia Hu, and James Caverlee. 2018. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1153–1162.

[220] Ziwei Zhu, Jingu Kim, Trung Nguyen, Aish Fenton, and James Caverlee. 2021. Fairness among New Items in Cold Start Recommender Systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 767–776.

[221] Ziwei Zhu, Jianling Wang, Yin Zhang, and James Caverlee. 2018. Fairness-aware recommendation of information curators. *The 2nd FATREC Workshop on Responsible Recommendation* (2018).