A Survey on the Fairness of Recommender Systems

YIFAN WANG, Tsinghua University, China
WEIZHI MA, Institute for AI Industry Research (AIR), Tsinghua University, China
MIN ZHANG*, Tsinghua University, China
YIQUN LIU, Tsinghua University, China
SHAOPING MA, Tsinghua University, China

Recommender systems are an essential tool to relieve the information overload challenge and play an important role in people's daily lives. Since recommendations involve allocations of social resources (e.g., job recommendation), an important issue is whether recommendations are fair. Unfair recommendations are not only unethical but also harm the long-term interests of the recommender system itself. As a result, fairness issues in recommender systems have recently attracted increasing attention. However, due to multiple complex resource allocation processes and various fairness definitions, the research on fairness in recommendation is scattered. To fill this gap, we review over 60 papers published in top conferences/journals, including TOIS, SIGIR, and WWW. First, we summarize fairness definitions in the recommendation and provide several views to classify fairness issues. Then, we review recommendation datasets and measurements in fairness studies and provide an elaborate taxonomy of fairness methods in the recommendation. Finally, we conclude this survey by outlining some promising future directions.

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

Additional Key Words and Phrases: recommendation, fairness, survey

1 INTRODUCTION

Nowadays, the amount of information available on the Internet has far exceeded individuals' information needs and processing capacity, which is known as information overload [40]. As a tool to alleviate information overload, recommender systems are widely used in people's daily lives (e.g., news recommendations, career recommendations, and even medical recommendations) and play a crucial role. Utility (such as click-through rate, dwell time, etc.) has been the most vital metric for recommender systems. However, only considering utility may lead to problems like the Matthew effect [111] and filter bubble [70]. Hence more views of recommender system performance have been proposed, such as diversity, efficiency, privacy, etc. Fairness is one of these critical issues.

* Corresponding author.
This work is supported by the Natural Science Foundation of China (Grant No. U21B2026, 62002191) and Tsinghua University Guoqiang Research Institute.
Authors' addresses: Yifan Wang, Min Zhang, Yiqun Liu, and Shaoping Ma, Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China. Emails: yf-wang21@mails.tsinghua.edu.cn, z-m@tsinghua.edu.cn, yiqunliu@tsinghua.edu.cn, msp@tsinghua.edu.cn; Weizhi Ma, Institute for AI Industry Research (AIR), Tsinghua University, Beijing 100084, China. Email: mawz@tsinghua.edu.cn.
Authors' addresses: Yifan Wang, Tsinghua University, Beijing, China, 10084, yf-wang21@mails.tsinghua.edu.cn; Weizhi Ma, Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China, 10084, mawz@tsinghua.edu.cn; Min Zhang*, Tsinghua University, Beijing, China, 10084, z-m@tsinghua.edu.cn; Yiqun Liu, Tsinghua University, Beijing, China, 10084, yiqunliu@tsinghua.edu.cn; Shaoping Ma, Tsinghua University, Beijing, China, 10084, msp@tsinghua.edu.cn.

1046-8188/2022/8-ART111
https://doi.org/10.1145/3547333

Recommender systems serve a resource allocation role in society by allocating information to users and exposure to items. Whether the allocation is fair can affect the personal experience and social good [73].

Fairness problems have received increasing attention from academia, industry, and society. Unfairness exists in different recommendation scenarios and various resources for both users and items. For users, there are significant differences in the recommendation accuracy between users of different ages and genders in movie recommendations and music recommendations, with female users and older users getting worse recommendation results [22]. In addition to accuracy, existing studies have also found considerable differences in other recommendation measurements such as diversity and novelty [91]. For items, existing research has found that minority items could get worse ranking performance and less exposure opportunity [6, 116]. Besides, in premium business scenarios, paid items may receive worse services from the platform than non-paid items [56]. Moreover, there are potential unfairness issues with various recommendation methods. Both traditional recommendation methods [22] and deep learning models [39] can suffer from unfairness.

Mitigating these unfairness phenomena is of great importance for recommender systems. Here are some reasons. (1) from an **ethical perspective**, as early as ancient Greece, fairness was listed by Aristotle as one of the crucial virtues to make people live well [3]. Fairness is an important virtue, and a fundamental requirement for a just society [74]. (2) from a **legal perspective**, Anti-discrimination laws [38] require that employment, admissions, housing, and public services do not discriminate against different groups of people based on gender, age, race, etc. For example, minority-owned companies should be recommended at a similar rate to white-owned companies in a job recommendation scenario [59]. (3) from a **user perspective**, a fair recommender system facilitates the exposure of different information in the recommendations, including some niche information, which may help break the information cocoon, alleviate the societal polarization, broaden users' horizons and enhance the value of recommendations. (4) from an **item perspective**, a fair recommender system can allocate more exposure to long-tail items, alleviating the Matthew effect [54]. It may also motivate these providers of niche items and then improve the diversity and creativity of items. (5) from a **system perspective**, a fair recommender system is conducive to its long-term interest. For example, an unfair recommender system may recommend popular content for users with niche interests, resulting in a bad experience. Similarly, it may also provide little exposure for niche providers. The lack of positive feedback may lead to a tendency for niche groups to leave the platform, which will reduce the diversity of content and users on the platform in the long run and affect the platform's growth [65]. Therefore, addressing unfairness is a critical issue for recommender systems.

A concept closely related to fairness is bias, which has also attracted extensive attention in current years. Some biases in recommender systems can lead to unfairness problems, such as popularity bias [113] and mainstream bias [53]. There are also some biases that have little to do with fairness, such as position bias. Generally speaking, fairness reflects normative ideas about how a recommender system should be, while bias is more concerned with statistical issues, such as the difference between what the model learns and the real world.

Although fairness has been studied in computer science for decades [27], and there is a lot of related work in machine learning [100, 101], fairness in recommendation has its unique problems. First, recommender systems are two-sided platforms that serve both users and items, where two-sided fairness needs to be guaranteed. Second, fairness in recommendation is dynamic in nature as there exists a feedback loop between users and the system. Third, on most platforms, the recommendation needs to be personalized by considering the unique needs of each user. Fairness in recommendation should also take users' personalization into account. Furthermore, apart from accuracy, fairness needs to be jointly considered with other measurements in the recommendation, such as diversity, explainability, and novelty. Therefore, current fairness work in machine learning, which mainly focuses on classification, could hardly be leveraged in recommender systems directly.

For the above reasons, fairness in recommendation has become an important topic in the research community. The attracted attention is increasing, which trends have been shown in Fig.1. As shown in Table 1, more than sixty fairness-related papers about recommendations have been published in top IR-related conferences and journals

(e.g., TOIS, SIGIR, WWW, and KDD) in recent five years. In the table, researches on fairness are summarized with their different definitions, targets, subjects, granularity, and optimization objects (details on these definitions are given in Table 3 and Section 3). We can find the focus of current studies. For example, consistent fairness (CO) is the most common definition of fairness, and current studies mainly focus on the group level. These trends are further discussed in the corresponding sections below.

Table 1. A lookup table for the reviewed papers about fairness in recommendation (Here "CO" means consistent fairness, "CA" means calibrated fairness, "CF" means counterfactual fairness, "EF" means envy-free fairness, "RMF" means Rawlsian maximin fairness, "PR" means process fairness and "MSF" means maximin-shared fairness, details on these definitions are given in Table 3).

Paper	Def.	Target		Subject			Granularity		Optim. Object		Pub.	Year
		Group	Individual	User	Item	Joint	Single	Amortized	Treatment	Impact		
[56]	CA	✓		✓	✓			✓		✓	BIGDATARES.	2022
[61]	CO	✓			✓			✓	✓		TOIS	2021
[93]	CO & CA	✓			✓		✓		✓	✓	WSDM	2021
[30]	CO	✓			✓			✓	✓		WSDM	2021
[36]	CO	✓			✓		✓		✓		WSDM	2021
[53]	CO	✓		✓				✓		✓	WSDM	2021
[95]	CO	✓		✓				✓	✓		AAAI	2021
[39]	CO	✓		✓				✓		✓	WWW	2021
[48]	CO & CA	✓			✓		✓		✓		WWW	2021
[96]	PR	-	-	✓			-	-	-	-	WWW	2021
[103]	CA	✓			✓			✓	✓		WWW	2021
[54]	CO	✓		✓				✓		✓	WWW	2021
[97]	CO & CA	✓	✓			✓		✓	✓		SIGIR	2021
[55]	CF		✓	✓			✓		✓		SIGIR	2021
[33]	CA	✓			✓			✓	✓	✓	SIGIR	2021
[115]	RMF		✓		✓			✓	✓		SIGIR	2021
[29]	RMF	✓	✓		✓		✓		✓		KDD	2021
[91]	CO	✓		✓				✓		✓	RECSYS	2021
[78]	CO		✓		✓			✓			CIKM	2021
[60]	CO		✓		✓			✓	✓		UMAP	2020
[83]	CO	✓			✓		✓		✓		UMAP	2020
[76]	CO	✓		✓				✓	✓		UMAP	2020
[20]	CO & CA	✓	✓		✓		✓		✓		CIKM	2020
[90]	CO & CA	✓				✓		✓	✓	✓	WSDM	2020
[71]	EF & MSF		✓			✓		✓	✓		WWW	2020
[107]	CO	✓			✓		✓		✓		WWW	2020
[45]	CO		✓	✓			✓		✓		RECSYS	2020
[59]	CA	✓			✓			✓	✓		PAKDD	2020
[112]	CO		✓		✓			✓	✓		ACCESS	2020
[28]	CO	✓	✓	✓				✓	✓		SIGIR	2020
[67]	CA	✓			✓			✓	✓	✓	SIGIR	2020
[116]	CO	✓			✓			✓	✓	✓	SIGIR	2020
[68]	CO	✓			✓		✓			✓	AAAI	2020
[86]	CO		✓	✓			✓		✓		SAC	2020
[19]	CA	✓		✓	✓			✓	✓	✓	RMSE	2019
[62]	CA	✓			✓		✓		✓		RMSE	2019
[9]	PR	-	-	✓	✓		-	-	-	-	ICML	2019

Table 1. (continued)

Paper	Def.	Target		Subject			Granularity		Optim. Object		Pub.	Year
		Group	Individual	User	Item	Joint	Single	Amortized	Treatment	Impact		
[31]	CA	✓			✓		✓		✓		KDD	2019
[6]	CO	✓			✓		✓			✓	KDD	2019
[73]	CO	✓	✓	✓	✓			✓		✓	WSDM	2019
[58]	CO	✓			✓		✓		✓		RECSYS	2019
[82]	CO	✓		✓				✓	✓		UMAP	2019
[94]	CO & CA	✓		✓				✓	✓		LOCALREC	2019
[8]	CO		✓		✓			✓	✓		MEDES	2019
[75]	CO		✓	✓			✓		✓		SAC	2019
[81]	CA		✓		✓		✓		✓		NIPS	2019
[25]	CO	✓		✓				✓		✓	FATREC	2018
[11]	-	✓		✓			-	-	-	-	FATREC	2018
[23]	CA	✓			✓			✓	✓		RECSYS	2018
[84]	CA	✓			✓		✓		✓		RECSYS	2018
[87]	CA	✓			✓			✓	✓		RECSYS	2018
[44]	CO	✓			✓		✓		✓		UMAP	2018
[52]	CO		✓	✓				✓	✓		WWW	2018
[85]	CO	✓			✓			✓	✓		WWW	2018
[7]	CA		✓		✓			✓	✓		SIGIR	2018
[64]	CO	✓			✓		✓		✓		CIKM	2018
[114]	CO	✓		✓	✓			✓	✓		CIKM	2018
[12]	CO	✓		✓	✓			✓	✓		FAT*	2018
[22]	CO	✓		✓				✓		✓	FAT*	2018
[43]	CO	✓		✓	✓			✓		✓	FAT*	2018
[15]	CA	✓			✓		✓		✓		ICALP	2018
[80]	CA	✓			✓		✓		✓		KDD	2018
[104]	CO	✓		✓				✓		✓	NIPS	2017
[77]	EF		✓	✓			✓		✓		WWW	2017
[99]	CO		✓	✓			✓		✓		RECSYS	2017
[66]	CA	✓			✓			✓	✓		PAKDD	2017
[106]	CO	✓			✓		✓		✓		CIKM	2017
[102]	CO	✓			✓		✓		✓		SSDBM	2017

Research on the fairness in recommendation is blossoming. However, due to various scenarios, diverse stakeholders, and different measurements, the research on fairness in the recommendation field is scattered. In order to fill this gap, this survey systematically reviews the existing research on fairness in the recommendation from several perspectives. The corresponding summary and discussion can guide and inspire future work. In summary, the contributions of this survey are as follows.

- We summarize existing **definitions** of fairness in recommendation and provide several **views** for classifying fairness issues in recommendation.
- We introduce some widely used **measurements** for fairness in recommendation and review fairness-related recommendation **datasets** in previous studies.
- We review current **methods** for fair recommendations and provide an elaborate taxonomy of methods.
- We outline several promising **future research directions** from the perspective of definition, evaluation, algorithm design, and explanation.

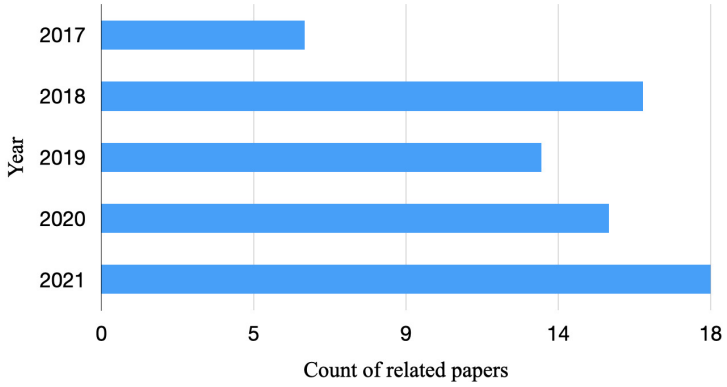


Fig. 1. The statistics of publications related to fairness in recommendation. We omit the work in 2022 in the figure, as most of the work in 2022 has not been published.

Several surveys are related to the topic of this survey. As far as we know, Castillo [13] firstly reviews fairness and transparency in information retrieval briefly. However, it only covers the related work before 2018, and in recent years, fairness in recommendations has developed greatly. [14, 63] concentrate on fairness in machine learning, but fairness in recommendation is not covered, especially its unique characteristics. Chen et al. [16] recently reviewed bias in recommender systems and introduced fairness issues, but fairness is not their main focus, and fairness measurements and datasets are not covered. To the best of our knowledge, there is no survey dedicated to systemically reviewing and detailing the fairness in the recommendation in a complete view.

This survey is structured as follows. In Section 2, we introduce existing definitions of fairness in the recommendation and discuss some related concepts. In Section 3, we present several perspectives to classify fairness issues in the recommendation. In Section 4, we introduce representative measurements for measuring fairness in the recommendation. In Section 5, we provide a taxonomy of methods to address unfairness in the recommendation. In Section 6, we introduce fairness-related datasets in recommender systems. In Section 7, we present possible future research directions. Finally, we conclude this survey in Section 8.

2 DEFINITIONS OF FAIRNESS IN RECOMMENDATION

In this section, we first provide definitions of fairness and then discuss the relationship between fairness and some related concepts in recommender systems. It is worth noting that discussions about fairness have existed since ancient times, but there is still no consensus on fairness. Due to the multitude of discussions related to fairness, it is impossible to list all relevant definitions. Therefore, we will introduce several definitions of fairness appearing in the research on recommendation, which can also be applied to other domains. The taxonomy of the reviewed fairness definitions is illustrated in Fig.2. To our knowledge, the definitions listed here are sufficient to cover the research on fairness in the recommendation. Besides, the notations used in the definitions are shown in Table 2.

2.1 Fairness Definitions

As we mentioned in the introduction, recommender systems play a resource allocation role in society, allocating information to users and exposure to items. For allocation, there are two aspects worthy of attention. One is the allocation process, such as the fairness of the recommendation model. The other is the allocation outcome,

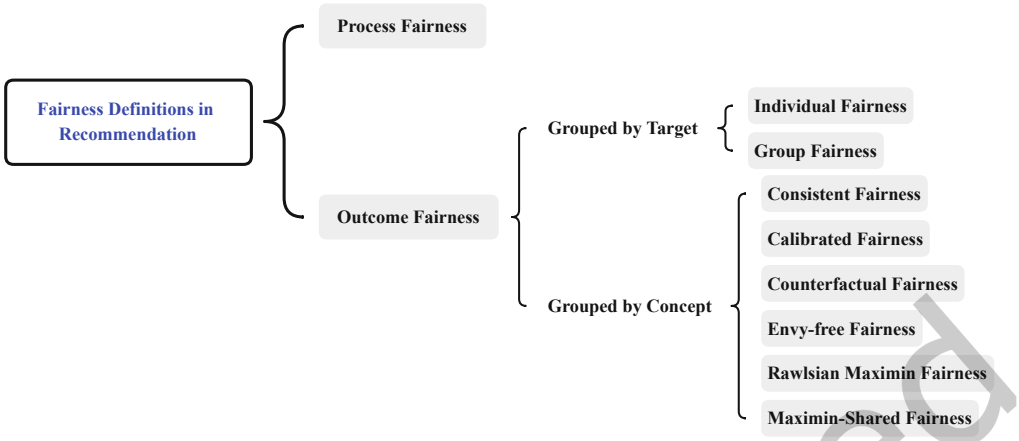


Fig. 2. Taxonomy of fairness definitions in recommendation.

Table 2. Notations used in fairness definitions and their explanations

Notation	Explanation
i	the i-th individual (e.g., a user or an item)
G_i	the i-th group of individuals
H	the hypothesis space of recommendation models
$C_h(\cdot)$	the outcome (e.g., predicted scores or recommendation lists) of model h given individuals or groups
$D(\cdot,\cdot)$	the distance function between individuals or groups
$D_o(\cdot,\cdot)$	the distance function between outcomes
$V(\cdot)$	the value function of outcomes
$V_i(\cdot)$	the personalized value function of outcomes for certain individual i
$M(\cdot)$	the merit function of individuals or groups

such as the fairness of the information received by users. Depending on whether the focus is on the process or outcome, fairness can be divided into process fairness and outcome fairness.

2.1.1 *Process Fairness.*

- ***Process Fairness.*** *Process fairness believes that the fair allocation should be fair in process [51, 69], which is also called procedural justice [51].*

Existing studies [34, 35, 51] generally focus on whether the information utilized in the allocation process is fair. In the case of job recommendation, process fairness concerns whether the recommendation model is fair, such as whether some unfair features (e.g., race) are used and whether the learned representations are fair.

2.1.2 *Outcome Fairness.*

- ***Outcome Fairness.*** *Outcome fairness holds that the fair allocation should lead to fair outcomes [24, 51], which is also called distributive justice [51].*

For example, in the case of job recommendation, outcome fairness concerns the recommendation outcome, such as whether whites would be more likely to be recommended than blacks even if they have the same ability.

The difference between these two kinds of fairness is similar to the difference between teleology and deontology in ethics. Teleology believes that whether a behavior is good or bad is related to outcomes, while deontology believes that it is only related to processes [4].

As the majority of existing research in recommendation focuses on the fairness of outcomes, we concentrate on definitions related to outcome fairness in the following. Outcome fairness can be further sub-grouped according to the target and concept.

Grouped by Target. Based on whether the target is to ensure group-level or individual-level fairness, outcome fairness can be further categorized into *group fairness* and *individual fairness*.

- ***Group Fairness.*** *Group fairness holds that outcomes should be fair among different groups.*

There are various ways to divide groups, and the most common is based on some explicit fairness-related attributes, such as gender, age, and race. When there are multiple fairness-related attributes, the whole may be divided into numerous subgroups. Fairness should be considered in these subgroups, as even if the groups under each single attribute division are fair, subgroups may be unfair to each other [46, 47].

- ***Individual Fairness.*** *Individual fairness believes that outcomes should be fair at the individual level.*

Individual fairness in some work refers to the idea that similar individuals should be treated similarly [7, 21]. However, there are also other definitions of individual-level fairness. For the sake of clarity, we use *individual fairness* to refer to a more general definition, i.e., fairness at the individual level.

Group fairness is more complex than individual fairness as different divisions may exist, and the divisions may be dynamic, i.e., one individual may belong to different groups at different times [30]. Moreover, individual fairness can be theoretically regarded as a special case of group fairness, in which each individual belongs to a unique group.

Grouped by Concept. Although fairness can be classified according to the target, we do not know what kind of outcomes are fair up to this point. About this, different researchers have different opinions, which we call fairness concepts. These concepts reflect researchers' understanding of what requirements should be met for fair outcomes. Compared to targets, fairness concepts include more information about fairness, and we can give more concrete formal definitions. We present these fairness concepts in the following.

A lot of papers [7, 30, 54] define fairness based on the similarity of the input (i.e., the individuals or groups receiving the allocation) and the output (i.e., the outcome of the allocation), which we call *consistent fairness*.

- ***Consistent Fairness.*** *At the individual level, consistent fairness argues that similar individuals should be treated similarly [7]. Formally, a fair model h should satisfy: for any two individuals i and j, if $D(i, j) \approx 0$, then $D_o(C_h(i), C_h(j)) \approx 0$. At the group level, consistent fairness requires that different groups should be treated similarly [30, 54]. Formally, a fair model h should satisfy: for any two groups G_i and G_j, it has $D_o(C_h(G_i), C_h(G_j)) \approx 0$.*

This concept of fairness first appeared in Aristotle's quote, "like cases should be treated alike" [3], which is thought to describe the consistency of fairness [42]. Dwork et al. [21] first formalized this definition at the individual level using the Lipschitz condition in a classification task. As a proper distance function of individuals is difficult to define, existing studies [28, 73] in recommendation usually use a trivial case as an alternative of consistent fairness, in which all individuals (or groups) are assigned similar outcomes. For the distance function of outcomes, current work often uses the difference between specific metrics (e.g., NDCG for users [22]) to measure distance.

- ***Calibrated Fairness.*** *Calibrated fairness [84] requires that the value of the outcome of an individual (or group) should be proportional to its merit, which is also called merit-based fairness [67]. Formally, at the individual level, a fair model h should satisfy: for any two individuals i and j, it has $\frac{V(C_h(i))}{M(i)} = \frac{V(C_h(j))}{M(j)}$. The group-level formalization is similar and only requires replacing i, j with G_i, G_j.*

This concept of fairness is closely related to Adams' Equity Theory [2]. Calibrated fairness requires two functions to measure the merit of individuals (or groups) and the value of the allocation outcome. The measure of merit often depends on the scenario, and the measure of value usually are some commonly used metrics (e.g., CTR for items [67]).

Most research focuses on the above two concepts of fairness, while a small number of papers explore other concepts.

- ***Envy-free Fairness.*** *Envy-free fairness requires that individuals should be free of envy, i.e., they should not be jealous of the results of others' outcomes [32]. Formally, a fair model h should satisfy: for every individual i and its outcome $C_h(i)$, it has $V_i(C_h(i)) \geq V_i(C_h(j))$ for any other individual j.*
- ***Counterfactual Fairness.*** *Counterfactual fairness requires that individuals have the same outcome in the real world as they do in the counterfactual world [72]. This means that if an individual belongs to a different group from the current one, its outcome will not change. Formally, a fair model h should satisfy: for every individual i, it has $C_h(i) = C_h(i)_{i \in G_j}$ for any other group G_j. The counterfactual $C_h(i)_{i \in G_j}$ can be calculated according to Pearl's three steps [72].*
- ***Rawlsian Maximin Fairness.*** *Rawlsian maximin fairness requires maximizing the value of the outcome of the worst individual or group [74]. Formally, at the individual level, a fair model h should satisfy: $h^* = \underset{h \in H}{argmax} \min_i V(C_h(i))$. The group-level formalization is similar and only requires replacing i with G_i.*
- ***Maximin-shared Fairness.*** *Maximin-shared fairness requires all individuals (or groups) to receive better outcomes than their maximin share [32]. Formally, a fair model h should satisfy: for every individual i and its outcome $C_h(i)$, given its personal value function $V_i(\cdot)$, we should have $V_i(C_h(i)) \geq MMS_i$, here $MMS_i = \max_{h \in H} \min_j V_i(C_h(j))$. The group-level formalization is similar and only requires replacing i, j with G_i, G_j.*

Table 3 divides the reviewed papers according to their definitions. It can be found that existing research pays more attention to outcome fairness. While in outcome fairness, previous studies mainly concentrate on group fairness in terms of target and focus on consistent fairness and calibrated fairness in terms of fairness concepts. Meanwhile, a few researchers have recently explored other concepts of fairness, such as Rawlsian maximin fairness.

Although many efforts about fairness definitions have been made, there are still some issues. Firstly, the relationships among these fairness definitions, especially in recommender systems, lack adequate exploration. If these fairness definitions conflict, which definition is more important is also a problem. Consensus on what kind of fairness should be achieved in recommender systems is necessary. Note that people may have different fairness needs [55], and consensus may not be the same in different scenarios. Besides, most studies concentrate on a single concept and target of fairness. Only a few recent studies [29, 71] attempt to achieve multiple fairness definitions simultaneously. If it is necessary to satisfy multiple fairness definitions, ensuring different fairness at the same time is also a question worth exploring.

2.2 Relationships between Fairness and Other Concepts

In this subsection, we discuss the relationship between fairness and some related concepts in recommender systems.

Bias. Bias is ubiquitous in recommender systems, which can exist in data, models, and outcomes [16]. Bias may increase both outcome unfairness and process unfairness. For example, Zhu et al. [113] demonstrate theoretically that matrix factorization models suffer from popularity bias in the learning process, which causes popular items to be preferred when the true preferences are the same. Besides, the inductive bias of representation learning may tend to learn some sensitive information to increase the information contained in the representation, which may

Table 3. A lookup table for the reviewed fairness definitions in recommendation.

Fairness Definitions		Abbr.	Description	Related Work
Process Fairness		PF	the allocation process should be fair	[9, 96]
Outcome Fairness		OF	the allocation outcome should be fair	all the below
Grouped by Target	Individual Fairness	IF	fairness should be guaranteed at the individual level	[28, 29, 45, 55, 60, 71, 73, 86, 97, 115] [7, 8, 20, 52, 75, 77, 78, 81, 99, 112]
	Group Fairness	GF	fairness should be guaranteed at the group level	[28–30, 39, 53, 54, 59, 83, 90, 93, 103] [12, 22, 23, 25, 43, 44, 58, 64, 73, 114] [20, 33, 36, 48, 76, 91, 95, 97, 102, 106] [6, 11, 15, 31, 56, 61, 66, 85, 104, 116] [19, 62, 67, 68, 80, 82, 84, 87, 94, 107]
Grouped by Concept	Consistent Fairness	CO	similar individuals / different groups should receive similar outcomes	[30, 36, 39, 48, 53, 54, 91, 93, 95, 97] [6, 28, 45, 68, 73, 86, 90, 107, 112, 116] [12, 22, 25, 43, 44, 52, 64, 85, 104, 114] [8, 20, 58, 60, 61, 75, 76, 78, 83, 94, 102] [82, 99, 106]
	Calibrated Fairness	CA	outcomes should be proportional to merits	[19, 20, 31, 33, 59, 62, 67, 81, 90, 103] [7, 15, 23, 56, 66, 80, 84, 87]
	Counterfactual Fairness	CF	individuals should have the same allocation outcome in the real world as they do in the counterfactual world	[55]
	Rawlsian Maximin Fairness	RMF	the outcomes of the worst should be maximized	[29, 115]
	Envy-free Fairness	EF	individuals should be free of envy	[71, 77]
	Maximin-Shared Fairness	MSF	individuals / groups should get better outcomes than their maximin share	[71]

increase process unfairness. Thus, removing fairness-related biases in data and models is helpful in alleviating unfairness [39]. Besides, there are also some biases that are not related to fairness, such as position bias. In general, bias is more concerned with statistical issues, while fairness all reflects normative ideas about how a recommender system should be.

Diversity. Diversity in recommendation means the diversity of items in the recommendation list, which is closely related to user satisfaction [26]. For item fairness, improvements in item fairness are possible to increase diversity. It is because when optimizing item fairness, the recommendation list tends to contain more cold items as well as items from more categories [44, 60], which means higher recommendation diversity. However, increasing diversity does not necessarily improve item fairness. The recommender system may recommend more popular items in each category, and cold items will still be treated unfairly. For user fairness, some studies find that existing methods to optimize recommendation diversity may exacerbate user unfairness [52]. Generally speaking, fairness is an evaluation criterion beyond diversity. Except for the fairness of accuracy, we can also consider the fairness of diversity [91].

Privacy. Privacy requires that external attackers cannot obtain sensitive information about users through recommendation results or parameters of the recommendation model [79]. Compared with privacy, fairness is an internal perspective of the recommender system, with no consideration of external attackers. Nevertheless, some fairness definitions may imply privacy, such as process fairness and counterfactual fairness. Process fairness requires that the recommendation process should be as fair as possible, such as using fair representations. If we consider that fair representations should be independent of fairness-related attributes, then a fair representation will also satisfy privacy for these attributes. Moreover, in the counterfactual perspective, Li et al. [55] demonstrate that counterfactual fairness of users can be guaranteed by making user representations independent of fairness-related attributes. This implies that user representations satisfying privacy can guarantee counterfactual fairness.

3 VIEWS OF FAIRNESS IN RECOMMENDATION

The definitions of fairness introduced in Section 2 can be applied to any allocation process and are not limited to the recommendation. Whereas, in recommender systems, there exist multiple allocation processes corresponding

to different fairness issues. In this section, to deepen the understanding of fairness, we present several views to classify fairness issues in the recommendation. These views and corresponding work are summarized in Table 4.

Table 4. A lookup table for the reviewed fairness work from several views.

Fairness Views		**Related Work**
Divided by Subject	User	[9, 19, 28, 39, 45, 53–55, 73, 76, 82, 86, 91, 95, 96] [11, 12, 22, 25, 43, 52, 56, 75, 77, 94, 99, 104, 114]
	Item	[20, 29, 30, 33, 36, 48, 59, 60, 78, 83, 93, 103, 107, 112, 115] [6, 8, 9, 19, 23, 31, 58, 62, 67, 68, 73, 81, 84, 87, 116] [7, 12, 15, 43, 44, 56, 61, 64, 66, 80, 85, 102, 106, 114]
	Joint	[71, 90, 97]
Divided by Granularity	Single	[6, 20, 29, 31, 36, 45, 48, 55, 58, 62, 68, 83, 86, 93, 107] [15, 44, 64, 75, 77, 80, 81, 84, 99, 102, 106]
	Amortized	[30, 33, 39, 53, 54, 60, 71, 76, 78, 90, 91, 95, 97, 103, 115] [8, 19, 23, 25, 28, 52, 59, 67, 73, 82, 85, 87, 94, 112, 116] [7, 12, 22, 43, 56, 61, 66, 104, 114]
Divided by Optimization Object	Treatment	[20, 29, 30, 33, 36, 48, 55, 60, 76, 83, 93, 95, 97, 103, 115] [19, 28, 31, 45, 58, 59, 62, 67, 71, 82, 86, 90, 107, 112, 116] [7, 8, 12, 15, 23, 44, 52, 64, 75, 81, 84, 85, 87, 94, 114] [61, 66, 77, 80, 99, 102, 106]
	Impact	[6, 19, 22, 25, 33, 39, 43, 53, 54, 56, 67, 68, 73, 90, 91, 93, 104, 116]

3.1 Subject

We refer to the subjects (e.g., individual i and group G_i in Section 2) receiving allocation in the allocation process as *fairness subject*, which corresponds to "Fair for Who." As there are different kinds of subjects in recommendation, fairness can be divided into **item fairness**, **user fairness**, and **joint fairness**. As demonstrated in Table 4, previous work mainly concentrates on item fairness and user fairness, and only a little work aims to improve joint fairness.

Item fairness concerns whether the recommendation treats items fairly, such as similar prediction errors for ratings of different types of items [73] or allocating exposure to each item proportional to its relevance [7]. If the recommendation treats different items unfairly, the providers of these discriminated items may lack positive feedback and leave the platform. Calibrated fairness is frequently applied to item fairness, while there is little work about calibrated fairness of users, probably because items are easily associated with concepts such as value and quality. The value of an item is often measured by its relevance to users [67] or the number of interactions in history [30]. Note that some researchers [1, 10] divide the subjects into consumer fairness and provider fairness. In contrast, we divide the subjects into user fairness and item fairness here, as provider fairness can be considered a kind of item fairness at the group level, where groups are divided according to providers.

User fairness concerns whether the recommendation is fair to different users, such as similar accuracy for different groups of users [22] or similar recommendation explainability across different users [28]. If the recommendation cannot be fair to users, it may lose users with specific interests. The most commonly used fairness definition in user fairness is consistent fairness, as it is often believed that different people are similar and should not be treated differently. However, here are some particular scenarios where fairness means treating people differently. For example, premium members should get better recommendations than standard members [19].

Moreover, there are some differences between user fairness in group recommendations and general recommendations. User fairness in general recommendations concerns all users [54], while group recommendations only care about the users in the group receiving the recommendation [77].

Joint fairness concerns whether both users and items are treated fairly [97]. In most recommendation scenarios, it is necessary to consider joint fairness, as user fairness and item fairness are vital to most recommender systems. It is worth noting that user fairness and item fairness can conflict with each other. When item fairness is improved, user fairness must worsen or remain the same [97], making joint fairness a challenging problem.

In addition to users and items, a few other stakeholders may exist in recommender systems. Their fairness issues have recently received attention from some researchers [1].

3.2 Granularity

We refer to the granularity of the allocation process as *fairness granularity*. Fairness in recommendation can be further divided into **single fairness** and **amortized fairness**.

A single recommendation list can be considered as the minimum allocation process in the recommendation, which corresponds to **single fairness**. Single fairness requires that the recommender system meets fairness requirements each time it generates a single recommendation list. In other words, the outcomes $C_h(\cdot)$ are only related to a single recommendation, and each recommendation should satisfy the specific fairness definition. For example, for item fairness, different types of items in a single recommendation list should satisfy the fair distribution [84]. For user fairness, a single recommendation list should be similarly relevant for different users in the group recommendation [77].

However, requiring every single recommendations list to be fair may be difficult and performance-damaging. An alternative is that we require the recommendations to be fair on the cumulative level, which is called **amortized fairness** [7]. Amortized fairness requires that the cumulative effect of multiple recommendation lists is fair, while a single recommendation list in them may be unfair. In other words, the outcomes $C_h(\cdot)$ are related to multiple recommendations.

For example, suppose we expect the exposure of books by male authors and books by female authors to be close in book recommendations. Single fairness requires that each recommendation list has approximately the same number of books by male authors as by female authors. In contrast, amortized fairness will only require that the system recommends approximately the same number of books by male authors as by female authors in all recommendations over time (e.g., within a day).

As demonstrated in Table 4, previous studies concentrate on amortized fairness, which is probably because single fairness is not achievable in some scenarios [7]. Existing work [67, 103] often uses the *average* value as the cumulative effect, such as the average exposure of a group across multiple recommendation lists. However, even if the *average* values are the same, the *variance* may be different, which may also be unfair. A high variance may mean that the recommendation performance is not stable and may bring more negative experiences to users. Nevertheless, no previous work has been done that takes *variance* into consideration.

3.3 Optimization Object

We refer to the aspect in which we are concerned about the allocation for subjects as *optimization object*, which is consistent with how the value function $V(\cdot)$ is defined in Section 2. There are many kinds of optimization objects, containing exposure and hit ratio of items [67], and accuracy of recommendations for users [91]. According to whether to consider the impact of allocation [5, 105], they can be divided into two main types, i.e., **treatment-based fairness** and **impact-based fairness**. Treatment-based fairness only considers whether the treatments of the recommender system are fair or not, such as the predicted scores to different users [114] and the allocated exposure to different items [71]. In contrast, impact-based fairness takes the impact caused by recommendations

(i.e., user feedback) into account. Taking item fairness as an example, in the Top-N ranking task, treatment-based fairness may require that the exposure of different items conforms to a fair distribution [31]. In contrast, impact-based fairness may require that the CTR of different items conforms to a fair distribution [67].

As shown in Table 4, most previous studies have focused on treatment-based fairness. It may be because it is more difficult to consider impact-based fairness as we cannot control user feedback directly. While most work only focuses on treatment-based fairness or impact-based fairness, it is also necessary to consider both impact-based fairness and treatment-based fairness together. Using item fairness as an example, on the one hand, if we only consider exposure without concerning the accuracy of recommendations, then there is a risk that the recommender system tends to recommend discriminated items to some inactive users. Although the exposure increases, the drop in click-through rate may instead lead to a loss of confidence of the provider. On the other hand, if we only consider the accuracy without considering the exposure of the recommendation, it may lead the recommender system to reduce the exposure chance of the discriminated items to reduce the decrease of the accuracy, which is also unfavorable for the discriminated items. Therefore, it is necessary to consider both impact-based fairness and treatment-based fairness.

4 MEASUREMENTS OF UNFAIRNESS IN RECOMMENDATION

4.1 Overview of Fairness Metrics

We introduce some widely used metrics for fairness in the recommendation, as shown in Table 5. Since there are different fairness definitions, the measurements of unfairness are not the same. Moreover, as the characteristics of fairness issues mentioned in Section 3 also affect the design and choice of fairness metrics, different metrics have different scopes of application, which are also marked in Table 5.

As demonstrated in Table 5, most fairness metrics are proposed for outcome fairness as it is the focus of most work, where more metrics for consistent fairness and calibrated fairness. Thus, we mainly present the corresponding metrics for these two fairness definitions in sections 4.2 and 4.3, respectively, and show all the others in section 4.4.

When selecting fairness metrics based on definitions, it is important to note that different metrics do not have the same scope of application. For consistent fairness, *Absolute Difference*, *Variance*, and *Gini coefficient* are commonly used measurements at the two-group, multi-group, and individual levels. These three metrics have a wide range of applicability to different subjects, granularity, and optimization objects. For calibrated fairness, *KL-divergence* and *L1-norm* are common measurements for multi-group and individual fairness. These two metrics also have broad applicability. Due to many groups in the group-level calibrated fairness studies, there are no metrics specifically designed for the two group situations. These common metrics are generic and can be used for both users and items but are relatively coarse-grained. They have two main drawbacks:

(1) These common metrics typically use the first-order moment like the average to describe groups, ignoring higher-order information;

(2) These metrics do not consider the characteristics of user fairness and item fairness.

In order to address the first point, some researchers [90, 114] use statistical tests like *KS statistic* or *ANOVA* that consider the population distribution. For the second point, for users, some researchers [104] consider user fairness on each item and then aggregate them. For items, some researchers [31, 102] consider unfairness across different positions and then aggregate them. Although limited in application, these metrics could be more proper for specific fairness issues. Specific details of these metrics are described below.

Since the metrics for different fairness definitions are not the same, we next present the corresponding metrics based on the fairness definitions. The meanings of the commonly used symbols are shown in Table 6.

Table 5. A lookup table for the reviewed fairness measurements with the order of Def. and the Target. "✓" denotes the presence of existing work using the metric under the corresponding conditions. "-" means that there is no work to use the metric in the corresponding condition, but the metric could theoretically be used in the corresponding condition as well. "×" indicates that the metric is not theoretically available for the corresponding condition. The abbreviations of the definitions are shown in Table 3. We use "1" to denote those measurements without a name in the original paper and "2" to denote those measurements with the original name.

Metric Name	Def.	Target			Subject		Granularity		Optim. Object		Related Work
		Two groups	More groups	Ind.	User	Item	Single	Amortized	Treat.	Impact	
Absolute Difference[1]	CO	✓	×	×	✓	✓	✓	✓	✓	✓	[28, 54, 93, 114]
KS statistic[2]	CO	✓	×	×	✓	-	-	✓	✓	-	[43, 114]
rND[2]	CO	✓	×	×	×	✓	✓	×	✓	×	[102]
rKL[2]	CO	✓	×	×	×	✓	✓	×	✓	×	[102]
rRD[2]	CO	✓	×	×	×	✓	✓	×	✓	×	[102]
Pairwise Ranking Accuracy Gap[2]	CO	✓	×	×	×	✓	✓	×	×	✓	[6, 93]
Value Unfairness[2]	CO	✓	×	×	✓	×	×	✓	×	✓	[25, 104]
Absolute Unfairness[2]	CO	✓	×	×	✓	×	×	✓	×	✓	[25, 39, 104]
Underestimation Unfairness[2]	CO	✓	×	×	✓	×	×	✓	×	✓	[25, 104]
Overestimation Unfairness[2]	CO	✓	×	×	✓	×	×	✓	×	✓	[25, 104]
Variance[2]	CO	-	✓	✓	✓	✓	✓	✓	✓	✓	[73, 97, 99]
Min-Max Difference[1]	CO	-	✓	✓	✓	✓	✓	✓	✓	-	[36, 86]
F-statistic of ANOVA[2]	CO	-	✓	×	✓	✓	-	✓	-	✓	[90]
Gini coefficient[2]	CO	-	-	✓	✓	✓	-	✓	✓	-	[28, 30, 52, 60, 61]
Jain's index[2]	CO	-	-	✓	✓	✓	✓	✓	✓	-	[99, 112]
Entropy[2]	CO	-	-	✓	-	✓	-	✓	✓	-	[60, 61, 71]
Min-Max Ratio[2]	CO	-	-	✓	✓	-	✓	-	✓	-	[45, 99]
Least Misery[2]	CO & RMF	-	-	✓	✓	-	✓	-	✓	-	[45, 75, 99]
MinSkew[2]	CA	-	✓	×	-	✓	✓	-	✓	-	[31]
MaxSkew[2]	CA	-	✓	×	-	✓	✓	-	✓	-	[31]
KL-divergence[2]	CA	-	✓	-	-	✓	✓	✓	✓	-	[56, 84, 90]
NDKL[2]	CA	-	✓	-	×	✓	✓	×	✓	-	[31]
JS-divergence[2]	CA	-	✓	-	-	✓	-	✓	✓	-	[66]
Overall Disparity[1]	CA	-	✓	-	-	✓	-	✓	✓	✓	[67, 103]
Generalized Cross Entropy[2]	CA	✓	✓	-	✓	✓	-	✓	✓	✓	[19, 56]
L1-norm[2]	CA	-	✓	✓	-	✓	✓	✓	✓	-	[7, 8, 48]
Proportion of Envy-free Users[1]	EF	×	×	✓	✓	×	✓	×	✓	-	[77]
Mean Average Envy[1]	EF	×	×	✓	✓	×	×	✓	✓	-	[71]
Classification-based Metrics[1]	CF & PR	×	×	✓	✓	✓	×	×	×	×	[9, 55, 96]
Bottom N Average[1]	RMF	×	-	✓	-	✓	-	✓	-	✓	[115]
Fraction of Satisfied Producer[1]	MSF	×	×	✓	×	✓	-	✓	✓	×	[71]

4.2 Metrics for Consistent Fairness (CO)

As mentioned in Section 2, current work on consistent fairness in recommendation requires that all individuals or groups should be treated similarly. Therefore, the corresponding measurements mainly measure the inconsistency of the utility distribution. Most metrics apply to both user fairness and item fairness. They consider the utility of each individual or group as a number and then measure the inconsistency of these numbers. Due to many metrics on consistent fairness and that early studies concentrate on situations where only two groups exist, we will present these metrics in the order of metrics for two groups, multiple groups, and individuals.

Absolute Difference. Absolute Difference (AD) is the absolute difference of the utility between the protected group G_0 and the unprotected group G_1. For user, the group utility $f(G)$ is often defined as the average predicted rating [114] or the average recommendation performance in the group G [28, 54]. For item, the group utility

Table 6. Notations and Explanations of Common Variables

Notation	Explanation
n	number of users
m	number of items
k	length of recommendation list
$\hat{r}_{u,i}$	prediction for user u and item i
$r_{u,i}$	feedback of user u to item i
$\mathcal{U} = \{u_1, ..., u_n\}$	the whole set of users
$\mathcal{I} = \{i_1, ..., i_m\}$	the whole set of items
$\mathcal{L} = \{l_{u_1}, ..., l_{u_n}\}$	the whole set of recommendation lists, $\|l_{u_i}\| = k$
$\mathcal{R} = \{r_{u,i}\}$	the whole set of feedback
$\mathcal{V}$	the whole set of individuals or groups, which can be either users or items
$f(\cdot)$	the utility function for individuals or groups

$f(G)$ can be defined as the whole exposure in the recommendation lists for the group G [93]. The lower the value, the fairer the recommendations.

$$AD = |f(G_0) - f(G_1)| \tag{1}$$

KS statistic. Kolmogorov-Smirnov statistic is a nonparametric test used to determine the equality of two distributions. It measures the area difference between two empirical cumulative distributions of the utilities for groups. The utilities are often defined as the predicted ratings in the group [43, 114]. Compared to *AD* using the average utility, KS statistic can measure the high-order inconsistency. The lower the value, the fairer the recommendations.

$$KS = |\sum_{i=1}^{T} l \times \frac{\mathcal{G}(R_0, i)}{|R_0|} - \sum_{i=1}^{T} l \times \frac{\mathcal{G}(R_1, i)}{|R_1|}| \tag{2}$$

Here T is the number of intervals in the empirical cumulative distribution, l is the size of each interval, $\mathcal{G}(R_0, i)$ is the number of utilities of the group G_0 that are inside the *i-th* interval.

rND, rKL and rRD. rND, rKL and rRD measure item exposure fairness for a ranking τ [102]. Unlike previous metrics, these metrics take the exposure position into account, calculating the normalized discounted cumulative unfairness similar to NDCG. Experiments show that rKD is smoother and more robust than rRD, and that rRD has limited application scope. The lower the value, the fairer the recommendations are for these metrics.

$$rND = \frac{1}{Z} \sum_{i=10,20,\dots}^{N} \frac{1}{\log_2 i} |\frac{|S^+_{1\dots i}|}{i} - \frac{|S^+|}{N}| \tag{3}$$

$$rKL = \frac{1}{Z} \sum_{i=10,20,\dots}^{N} \frac{1}{\log_2 i} (\frac{|S^+_{1\dots i}|}{i} \log \frac{\frac{|S^+_{1\dots i}|}{i}}{\frac{|S^+|}{N}} + \frac{|S^-_{1\dots i}|}{i} \log \frac{\frac{|S^-_{1\dots i}|}{i}}{\frac{|S^-|}{N}}) \tag{4}$$

$$rRD = \frac{1}{Z} \sum_{i=10,20,\dots}^{N} \frac{1}{\log_2 i} |\frac{|S^+_{1\dots i}|}{|S^-_{1\dots i}|} - \frac{|S^+|}{|S^-|}| \tag{5}$$

Here the normalizer Z is the highest possible value of corresponding measurements, $|S^+_{1\dots i}|$ is the number of the protected group in the top-i of the ranking τ, S^+ is the number of the unprotected group in the whole ranking.

Pairwise Ranking Accuracy Gap. Pairwise Ranking Accuracy Gap (PRAG) measures item unfairness in the pairwise manner [6, 93]. Unlike previous metrics focusing on exposure or click-through rate, PRAG measures the

unfairness of pairwise ranking accuracy, and it is calculated on data from randomized experiments. The lower the value, the fairer the recommendations.

$$PRAG = |PairAcc(I_1 > I_2|q) - PairAcc(I_1 < I_2|q)| \tag{6}$$

$$PairAcc(I_1 > I_2|q) = P(f(x_i) > f(x_j)|y_i > y_j, i \in I_1, j \in I_2) \tag{7}$$

Here $PairAcc$ represents the ranking accuracy for a pair of items x_i, x_j from different groups I_1, I_2. $f(x_i)$ and $f(x_j)$ are the predicted score for the recommendation query q. y_i and y_j are the true feedback, which are collected through randomized experiments.

Value Unfairness and its variants. Value unfairness is proposed to measure inconsistency in signed prediction error between two user groups [104]. There are three variants of Value unfairness. Absolute Unfairness measures the inconsistency of absolute prediction error, while Underestimation Unfairness and Overestimation Unfairness measure inconsistency in how much the predictions underestimate and overestimate the true ratings, respectively. The lower the value, the fairer the recommendations.

$$U_{val} = \frac{1}{m}\sum_{i=1}^{m} |(E_0[\hat{r}]_i - E_0[r]_i) - (E_1[\hat{r}]_i - E_1[r]_i)| \tag{8}$$

$$U_{abs} = \frac{1}{m}\sum_{i=1}^{m} ||E_0[\hat{r}]_i - E_0[r]_i| - |E_1[\hat{r}]_i - E_1[r]_i|| \tag{9}$$

$$U_{under} = \frac{1}{m}\sum_{i=1}^{m} |max(0, E_0[r]_i - E_0[\hat{r}]_i) - max(0, E_1[r]_i - E_1[\hat{r}]_i)| \tag{10}$$

$$U_{over} = \frac{1}{m}\sum_{i=1}^{m} |max(0, E_0[\hat{r}]_i - E_0[r]_i) - max(0, E_1[\hat{r}]_i - E_1[r]_i)| \tag{11}$$

Here $E_0[\hat{r}]_i$ is the average predicted score for the $i\text{-}th$ item from group 0, and $E_0[r]_i$ is the average rating for the $i\text{-}th$ item from group 0.

The above metrics are only applicable to measure inconsistency between two groups. In the following, we present the metrics to measure unfairness for **three or more groups**. It is worth noting that since we can consider individual fairness as a special case of group fairness (i.e., each individual belongs to a unique group), theoretically, these group fairness metrics below can also apply to individual fairness. However, in practice, the common metrics for individual and group fairness are different.

Variance. Variance is a commonly used metrics for dispersion, which is applied to both group-level [73, 97] and individual-level [73, 97, 99]. The utility can be the rating prediction error [73], the predicted recommendation satisfaction for a single user [97, 99] and the average exposure for an item group [97]. The lower the value, the fairer the recommendations.

$$Variance = \frac{1}{|\mathcal{V}|^2}\sum_{v_x \neq v_y} (f(v_x) - f(v_y))^2 \tag{12}$$

Min-Max Difference. Min-Max Difference (MMD) is the difference between the maximum and the minimum of all allocated utilities. This metric is used to measure the inconsistency of the average exposure for multiple item groups [36], and the disagreement for users in group recommendation at the individual level [86]. The lower the value, the fairer the recommendations.

$$MMD = max\{f(v), \forall v \in \mathcal{V}\} - min\{f(v), \forall v \in \mathcal{V}\} \tag{13}$$

F-statistic of ANOVA. The one-way analysis of variance (ANOVA) is used to determine any statistically significant differences between the mean values of three or more independent groups. Its F-statistic can be

considered a fairness measurement. The utility can be the rating prediction error for a single rating [90]. The lower the value, the fairer the recommendations.

$$F = \frac{MST}{MSE} \tag{14}$$

$$MST = \frac{\sum_i |v_i| \times (\overline{v}_i - \overline{v})^2}{|\mathcal{V}| - 1} \tag{15}$$

$$MSE = \frac{\sum_i \sum_{j \in v_i} (f(ind_j) - \overline{v}_i)^2}{\sum_{v \in \mathcal{V}} |v| - |\mathcal{V}|} \tag{16}$$

Here $f(ind_j)$ is the utility of an individual belong to v_i, $\overline{v}_i$ is the mean utility of group v_i, $\overline{v}$ is the mean utility of all individuals.

In the following, we present some metrics commonly used for **individual fairness**. Note that in addition to the metrics below, $Variance$ above is also often used to measure individual fairness.

Gini coefficient. Gini coefficient is widely used in sociology and economics to measure the degree of social unfairness [28, 30, 52, 60, 61]. To our knowledge, it is also the most commonly used metric for consistent individual fairness. The utility can be the predicted relevance for a user [28, 52] or the exposure for an item [30, 60, 61]. The lower the value, the fairer the recommendations.

$$Gini = \frac{\sum_{v_x, v_y \in \mathcal{V}} |f(v_x) - f(v_y)|}{2|\mathcal{V}| \sum_v f(v)} \tag{17}$$

Jain's index. Jain's index [41] is commonly used to measure unfairness in network engineering. Some studies use it to measure the inconsistency of predicted user satisfaction in group recommendations [99] and the inconsistency of item exposure [112]. The higher the value, the fairer the recommendations.

$$Jain = \frac{(\sum_v f(v))^2}{|\mathcal{V}| \cdot \sum_v f(v)^2} \tag{18}$$

Entropy. Entropy is often used to measure the uncertainty of a system. In recommendation, it is used to measure the inconsistency of item exposure [60, 61, 71]. The lower the value, the fairer the recommendations.

$$Entropy = -\sum_{v \in \mathcal{V}} p(v) \cdot \log p(v) \tag{19}$$

Min-Max Ratio. Min-Max Ratio is the ratio of the minimum to the maximum of all allocated utility. Some studies [45, 99] use it to measure the inconsistency of the predicted user satisfaction in group recommendation. The higher the value, the fairer the recommendations.

$$MinMaxRatio = \frac{min\{f(v), \forall v \in \mathcal{V}\}}{max\{f(v), \forall v \in \mathcal{V}\}} \tag{20}$$

Least Misery. Least Misery is the minimum of all allocated utility. It is also a commonly used fairness metric in group recommendation [45, 75, 99]. The higher the value, the fairer the recommendations.

$$LeastMisery = min\{f(v), \forall v \in \mathcal{V}\} \tag{21}$$

4.3 Metrics for Calibrated Fairness (CA)

Calibrated fairness requires defining the merit of an individual or group. We denote $Merit(\cdot)$ as a merit function that measures the merit of an individual or group. We can calculate the fair distribution of the allocation based on $Merit(\cdot)$, i.e., the proportion of the individual's or group's allocation to the total allocation in the fair case, i.e., $p_f(v_i) = \frac{Merit(v_i)}{\sum_j Merit(v_j)}$. We can also calculate the proportion of the total allocation for an individual or group in the current situation, i.e., $p(v_i) = \frac{f(v_i)}{\sum_j f(v_j)}$. Most measurements of calibrated fairness measure the difference between the distribution of utilities p and the distribution of merits p_f.

Since all the group fairness metrics in calibrated fairness can be applied to multiple groups, we will present them in the order of group fairness and individual fairness.

***MinSkew** and **MaxSkew**.* The deviation (Skew) on a certain group v can be defined as $\log(\frac{p_f(v)}{p(v)})$. And then, we can define the min-skew and the max-skew as follows. Here the utility can be the exposure of the item group, while the p_f is a predefined distribution [31]. For MinSkew, the higher the value, the fairer the recommendations. For MaxSkew, the lower the value, the fairer the recommendations.

$$Min - Skew = min\{\log(\frac{p_f(v)}{p(v)}), v \in \mathcal{V}\} \tag{22}$$

$$Max - Skew = max\{\log(\frac{p_f(v)}{p(v)}), v \in \mathcal{V}\} \tag{23}$$

KL-divergence. KL-divergence measures how one probability distribution is different from the other. It can be used to measure the difference between p_f and p. Here the utility can be the exposure of the item group, while the p_f can be calculated by the group's historical exposure [56, 84, 90]. The lower the value, the fairer the recommendations.

$$D_{KL}(p, p_f) = \sum_{v \in \mathcal{V}} p(v) \log \frac{p(v)}{p_f(v)} \tag{24}$$

NDKL. NDKL is an item unfairness measure based on KL-divergence [31]. It computes the KL-divergence for each position and then obtains a normalized discounted cumulative value. The lower the value, the fairer the recommendations.

$$NDKL@K = \frac{1}{Z} \sum_{i}^{K} \frac{1}{\log(i+1)} D^i_{KL} \tag{25}$$

Here the normalizer Z is computed as the highest possible value, and D^i_{KL} is the KL-divergence of the top-i ranking.

JS. Like KL-divergence, JS-divergence also measures how one probability distribution differs from the other. Some work [66] uses JS-divergence as a metric instead of KL-divergence as it is symmetrical while KL-divergence is asymmetrical. The lower the value, the fairer the recommendations.

$$D_{JS}(p, p_f) = \frac{1}{2}(D_{KL}(p, \frac{1}{2}(p_f + p)) + D_{KL}(p_f, \frac{1}{2}(p_f + p))) \tag{26}$$

Overall Disparity. Overall disparity measures the average disparity of the proportion of the utility and merit among different groups. The utility can be exposure-based or click-based [67, 103]. The lower the value, the fairer the recommendations.

$$OD = \frac{2}{|V|(|V|-1)} \sum_{i=0}^{|V|} \sum_{j=i+1}^{|V|} || \frac{p(v_i)}{p_f(v_i)} - \frac{p(v_j)}{p_f(v_j)} || \tag{27}$$

Generalized Cross Entropy. Generalized cross entropy [19, 56] also measures how one probability distribution is different from the other. The higher the value, the fairer the recommendations.

$$GCE = \frac{1}{\alpha(1-\alpha)}\left[\sum_{v\in\mathcal{V}} p_f^{\alpha}(v)p^{(1-\alpha)}(v) - 1\right] \tag{28}$$

Here α is a hyperparameter.

In the following, we present calibrated fairness measures frequently used at the individual level.

L1-norm. L1-norm is the sum of the magnitudes of the vectors in a space. Some researchers [7, 8, 48] treat the merit and utility distributions as vectors and then use the L1-norm to calculate the distance between the vectors. This metric is often used for individual-level measurement [7, 8], and there is also work [48] that uses it to measure group-level unfairness. The lower the value, the fairer the recommendations.

$$L1-norm = \sum_{v\in\mathcal{V}} |p(v) - p_f(v)| \tag{29}$$

It is worth noting that some measures of calibrated fairness and consistent fairness are interconvertible. Theoretically, for a calibrated fairness measurement, if we set p_f to a uniform distribution, it can become a measurement for consistent fairness. On the other hand, for a consistent fairness measurement which contains $f(v)$, we can set $f(v)$ to $\frac{p(v)}{p_f(v)}$, then it become a calibrated fairness measurement.

4.4 Metrics for Other Fairness Definitions

4.4.1 Metrics for Envy-free Fairness (EF). Envy-free fairness requires a definition of envy, which can be different in different scenarios. In group recommendations, different users in the group receive the same recommendations. Serbos [77] defines envy as follow:

Envy-freeness (in group recommendation). Given a group G, a group recommendation package P, and a parameter δ, we say that a user $u \in G$ is envy-free for an item $i \in P$ if $r_{u,i}$ is in the top-δ% of the preferences in the set $\{r_{v,i} : v \in G\}$.

This envy definition can be applied to a single item. This definition means that a user u feels envy on an item if at least δ% users in the group like this item more than u. It is impossible for all users in a group to be envy-free (i.e., the user is envy-free for all items in the package). In practice, m-envy-free is often used, which means that the user in the group is envy-free for at least m items.

A measurement for envy-free fairness can be the proportion of m-envy-free users:

$$F = \frac{|G_{ef}|}{|G|} \tag{30}$$

where $|G_{ef}|$ is the number of m-envy-free users. The higher the value, the fairer the recommendations.

In general recommendations, different users receive different recommendations. Patro et al. [71] define envy-freeness as follow:

Envy-freeness(in general recommendation). Given a utility metrics f and all the recommendation lists $\mathcal{L}$, we say that a user u is envy-free for a user v if and only if $f(l_v, u) \geq f(l_u, u)$ and the degree of envy can be defined as $max(f(l_v, u) - f(l_u, u), 0)$. Here $f(l, u)$ is the predicted relevance sum for the user u with the recommendation list l.

This envy definition is applied to each pair of users. Unlike envy in group recommendations, this definition does not involve the third user. Moreover, it is feasible to make all users envy-free with utility metrics properly chosen.

The average of envy among users can be a measurement of envy-free fairness:

$$Envy(\mathcal{U}) = \frac{1}{n \cdot (n-1)} \sum_{u_i, u_j, u_i \neq u_j} envy(u_i, u_j) \tag{31}$$

where $envy(u_i, u_j) = max(f(l_i, u_i) - f(l_j, u_i), 0)$. The lower the value, the fairer the recommendations.

4.4.2 Metrics for Counterfactual Fairness (CF). Li et al. [55] demonstrate that counterfactual user fairness can be guaranteed when user embeddings are independent of fairness-related attributes. Therefore, they use a classifier to predict fairness-related attributes based on user embeddings and use classification measurements to measure counterfactual fairness. The classification measurements can be Precision, Recall, AUC, and F1 et al.

4.4.3 Metrics for Rawlsian Maximin Fairness (RMF). Rawlsian maximin fairness argues that fairness depends on the worst individual or group. A simple measurement is the utility of the worst case, but it is vulnerable to noise. In order to make the metrics robust, some work [115] uses the average utility of the bottom n% as a measurement. The higher the value, the fairer the recommendations.

4.4.4 Metrics for Maximin-shared Fairness (MSF). Maximin-shared fairness requires the outcome of each individual to be more than its maximin share. A measurement for item maximin-shared fairness is the proportion of individuals satisfying this condition, where the maximin share for every item is a constant value, i.e., the average exposure[71]. The higher the value, the fairer the recommendations.

4.4.5 Metrics for Process Fairness (PR). One criterion of process fairness is that the model should use fair representations. A fair representation should be independent of fairness-related attributes, so some work [9, 96] trains a classifier to predict fairness-related attributes of users and items according to their representations. Then they use some classification measurements (e.g., precision) to measure the fairness of representations, which are similar to the counterfactual fairness measurements [55].

5 METHODS FOR FAIR RECOMMENDATION

5.1 Overview of Fairness Methods

To our knowledge, existing methods for improving fairness can be categorized into three classes according to their working position in the recommendation pipeline, i.e., data-oriented methods, ranking methods, and re-ranking methods, as shown in Fig.3. Data-oriented methods are proposed to alleviate the unfairness problem by changing the training data. Ranking methods mainly design fairness-aware recommendation models or optimization targets for learning fair recommendations. Re-ranking methods mainly adjust the outputs of recommendation models to improve fairness. Since there are more methods in the last two categories, we further grouped the methods in these two categories, and the specific sub-groups are also illustrated in Fig.3.

The reviewed methods and corresponding brief descriptions are summarized in Table 7. It can be observed that there are only a few data-oriented methods. For ranking methods, regularization and adversarial learning are the dominant methods. At the same time, reinforcement learning has also gained attention in recent years due to being more suitable for modeling dynamics and long-term effects. For re-ranking methods, slot-wise re-ranking methods are dominant, but an increasing amount of recent work has focused on global-wise re-ranking.

As shown in Table 8, we also summarize the types of fairness issues solved by each method type. It can be found that each method type can solve several different types of fairness issues, and most fairness issues, on the other hand, can be solved using multiple methods. However, some fairness issues are more specific. For example, process fairness and counterfactual fairness issues are solved only using adversarial learning. Rawlsian maximin and maximin-shared fairness tend to be solved using global-wise re-ranking. Indeed, it may be because there is less work related to these fairness issues. It is also worth exploring to design other methods to solve these issues.

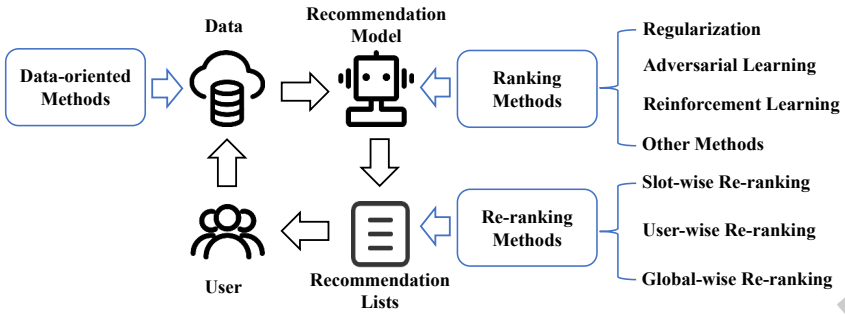


Fig. 3. Taxonomy of fairness methods in the recommendation and their position in the recommendation pipeline.

5.2 Data-oriented Methods

The data-oriented methods improve fairness by modifying the training data. Compared with other types of methods, there are fewer data-oriented methods.

Considering that user unfairness might result from the data imbalance between different user groups, Ekstrand et al. [22] use re-sampling to adjust the proportion of different user groups in the training data. Experiments on the Movielens 1M dataset show that this approach can alleviate unfairness, but not significantly.

Rastegarpanah et al. [73] design a relatively more complex but effective method. They draw on data poisoning attacks to address the unfairness problem by adding additional antidote data (e.g., fake user data) to the training data. Adding antidote data during training will affect the predicted rating matrix, which further affects the fairness of recommendations. The antidote data can be updated by optimizing the fairness objective function through the gradient descent method. Compared to the re-sampling method, this approach can better mitigate unfairness, but it is also relatively more time-consuming.

In summary, we can adjust the training data to improve the fairness of recommendations. The advantage of these methods is their low coupling with the recommender system since these methods do not require modification of the original recommendation model. Besides, as these methods work at the front part of the recommendation pipeline, there are fewer constraints on the candidate set. They have the potential to improve the fairness of the recommendation results significantly. However, since multiple stages exist between the data and the final presentation, their performance might be degraded by subsequent stages such as re-ranking for diversity. It is challenging to design effective data-oriented methods.

5.3 Ranking Methods

Ranking methods mainly modify recommendation models or optimization targets to learn fair representations or prediction scores. The ranking is the main focus of research in recommendation techniques. It is natural to use some advanced techniques to solve the problems of fair representation learning and long-term fairness, which is difficult for the other two types of methods. Compared to data methods, the results of sorting methods are less different from the final presentation, and the improvement in fairness is more straightforward. Nevertheless, since a re-ranking stage may exist after the ranking stage, similar to data-oriented methods, their performance may be damaged by downstream re-ranking stages.

Depending on the different techniques, current fairness methods for the ranking phase can be divided into regularization-based methods, adversarial learning-based methods, reinforcement learning-based methods, and others.

Table 7. A lookup table for the reviewed fairness methods in Recommendation.

Paper	Type	Brief Description	Publication	Year
[22]	data-oriented	adjust the proportion of the protected group by resampling	FAT*	2018
[73]	data-oriented	add antidote data to the training data	WSDM	2019
[104]	regularization	use fairness metrics (e.g., value fairness) as fair regularization	NIPS	2017
[43]	regularization	use distribution matching and mutual information terms as regularization	FAT*	2018
[12]	regularization	add fairness regularization to SLIM	FAT*	2018
[114]	regularization	induce orthogonality between insensitive latent factors and sensitive factors	CIKM	2018
[6]	regularization	add pairwise fairness regularization based on randomized experiments	KDD	2019
[90]	regularization	use F-statistic of ANOVA as regularization	WSDM	2020
[9]	adversarial learning	fairness constraints for graph embeddings	ICML	2019
[116]	adversarial learning	learn fair predicted scores by enhancing score distribution similarity	SIGIR	2020
[96]	adversarial learning	learn fair representations in graph-based recommendation	WWW	2021
[53]	adversarial learning	add text-based reconstruction loss to learn fair representations	WSDM	2021
[55]	adversarial learning	learn personalized counterfactual fair user representations	SIGIR	2021
[95]	adversarial learning	learn fair user representation in news recommendation	AAAI	2021
[59]	reinforcement learning	add fairness-related rewards to improve long term fairness	PAKDD	2020
[30]	reinforcement learning	add fairness-related constraints to improve long term fairness	WSDM	2021
[25]	other ranking method	a hybrid fair model with probabilistic soft logic	RECSYS	2018
[8]	other ranking method	add a noise component to VAE	MEDES	2019
[39]	other ranking method	use a pre-training and fine-tuning approach with bias correction techniques	WWW	2021
[56]	other ranking method	adjust the gradient based on the predefined fair distribution	BIGDATARES.	2022
[106]	slot-wise re-ranking	maximize ranking utility with group fairness constraint by two queues	CIKM	2017
[77]	slot-wise re-ranking	use greedy algorithm to maximize fairness in group recommendation	WWW	2017
[44]	slot-wise re-ranking	fairness-aware variation of the maximal marginal relevance	UMAP	2018
[84]	slot-wise re-ranking	calibrated recommendation through maximal marginal relevance	RECSYS	2018
[31]	slot-wise re-ranking	improve multiple group fairness by interval constrained sorting	KDD	2019
[75]	slot-wise re-ranking	find pareto optimal items in group recommendation	SAC	2019
[58]	slot-wise re-ranking	personalized fairness-aware re-ranking	RECSYS	2019
[83]	slot-wise re-ranking	personalized fairness-aware re-ranking with different user tolerance	UMAP	2020
[45]	slot-wise re-ranking	ensure fairness in group recommendation in a ranking sensitive way	RECSYS	2020
[67]	slot-wise re-ranking	ensure fairness in dynamic learning to rank through p-controller	SIGIR	2020
[103]	slot-wise re-ranking	ensure fairness in dynamic learning to rank by maximal marginal relevance	WWW	2021
[99]	user-wise re-ranking	fair group recommendation from the perspective of Pareto Efficiency	RECSYS	2017
[64]	user-wise re-ranking	a series of recommendation policies to combine fairness and relevance	CIKM	2018
[7]	user-wise re-ranking	ensure amortized fairness through integer linear programming	SIGIR	2018
[80]	user-wise re-ranking	linear programming from the perspective of probabilistic rankings	KDD	2018
[87]	global-wise re-ranking	0-1 integer programming with providers constraint	RECSYS	2018
[71]	global-wise re-ranking	a re-ranking method for both user fairness and item fairness	WWW	2020
[28]	global-wise re-ranking	fairness-aware explainable recommendation through 0-1 integer programming	SIGIR	2020
[61]	global-wise re-ranking	a re-ranking method based on maximum flow	TOIS	2021
[54]	global-wise re-ranking	ensure user group fairness through 0-1 integer programming	WWW	2021
[97]	global-wise re-ranking	a re-ranking method for both user fairness and provider fairness	SIGIR	2021
[115]	global-wise re-ranking	a learnable re-ranking method for fairness among new items	SIGIR	2021

5.3.1 Regularization. One common approach is adding a fairness-related regularization term to the loss function. Formally, denote L_{rec} as the traditional recommendation loss function and L_{fair} as the fairness-related regularization term, then the loss function considering fairness is formalized as $L = L_{rec} + \lambda \cdot L_{fair}$.

One **direct** approach is to add the fairness evaluation metrics [43, 90, 104, 107] to the loss function as a regularization term, which requires that the metric is differential. It is difficult to use this approach to address unfairness in exposure or ranking as the corresponding metrics are not differential, so existing related work is more

Table 8. The current types of fairness issues solved by each method type. (Here "CO" means consistent fairness, "CA" means calibrated fairness, "CF" means counterfactual fairness, "EF" means envy-free fairness, "RMF" means Rawlsian maximin fairness, "PR" means process fairness and "MSF" means maximin-shared fairness).

Method Type	Def.	Target		Subject			Granularity		Optim. Object	
		Group	Ind.	User	Item	Joint	Single	Amortized	Treat.	Impact
Data-oriented [22, 73]	CO	✓	✓	✓	✓			✓		✓
Regularization [6, 12, 43, 90, 104, 114]	CO & CA	✓		✓	✓	✓	✓	✓	✓	✓
Adversarial Learning [9, 53, 55, 95, 96, 116]	PR & CO & CF	✓	✓	✓	✓		✓	✓	✓	✓
Reinforcement Learning [30, 59]	CO & CA	✓			✓			✓	✓	
Others [8, 25, 39, 56]	CO & CA	✓	✓	✓	✓			✓	✓	✓
Slot-wise Re-ranking [31, 44, 58, 75, 77, 84, 106] [45, 67, 83, 103]	CO & CA & EF	✓	✓	✓	✓		✓	✓	✓	✓
User-wise Re-ranking [7, 64, 80, 99]	CO & CA	✓	✓	✓	✓		✓	✓	✓	✓
Global-wise Re-ranking [28, 54, 61, 71, 87, 97, 115]	CO & CA & EF & MSF & RMF	✓	✓	✓	✓	✓		✓	✓	✓

focused on unfairness in rating prediction. The advantage of this approach is its simplicity and effectiveness, while the disadvantage is that it is limited in application and often results in a loss of recommendation performance.

In contrast, some approaches [6, 12, 114] impose **indirect** regularization on the model. Compared to direct methods, indirect methods can achieve better fairness and recommendation performance. Here we introduce some representative methods below.

In order to reduce the correlation between predicted scores and fairness-related attributes, Zhu et al. [114] propose a fairness-aware tensor-based recommendation framework(FATR), which induces orthogonality between the representations of users (or items) and the corresponding vector of fairness-related attributes by adding a regular term in the tensor-based recommendation model. The loss function is the following Eq.(32) [114]. The first term of the loss function is the original part of the tensor-based recommendation model. The second term is the regular fairness term, which extracts fairness-related attribute information in latent factor matrices. The final fairness prediction is calculated as $[\![A'_1, ..., A'_n, ..., A'_N]\!]$.

$$\begin{aligned} \underset{X, A_1, \ldots, A'_n, \ldots, A_N}{minimize} \quad & L = ||X - [\![A_1, ..., \tilde{A_n}, ..., A_N]\!]||_F^2 + \frac{\lambda}{2}||A_n''^T A_n'||_F^2 + \frac{\gamma}{2}\sum_{i=1}^{N}||A_i||_F^2 \\ s.t. \quad & \Omega \circledast X = J, \tilde{A_n} = [A'_n \; A''_n], A''_n = S \end{aligned} \tag{32}$$

Here X is a tensor denoting the complete preferences of users, $[\![\cdot]\!]$ is the Kruskal operator, $[A\ B]$ is the matrices concatenating operator, $\odot$ is the Khatri-Rao product, and $\circledast$ is the Hadamard product. J denotes observations, and Ω is the non-negative indicator tensor indicating whether we observe X. $A_1, ..., A_N$ denote the latent factor matrices of all the modes of the tensor. Here $A_n \in R^{d_n \times r}$ is the latent factor matrix of the fairness-related mode

mode-n, where r is the dimension of the latent factors and d_n is the entity number of the mode-n, and it can be split into two part A'_n, A''_n.

Experiments on real datasets show that FATR can achieve better recommendation performance and fairness than directly using the fairness metric as a regular term, reflecting the advantages of the indirect approach.

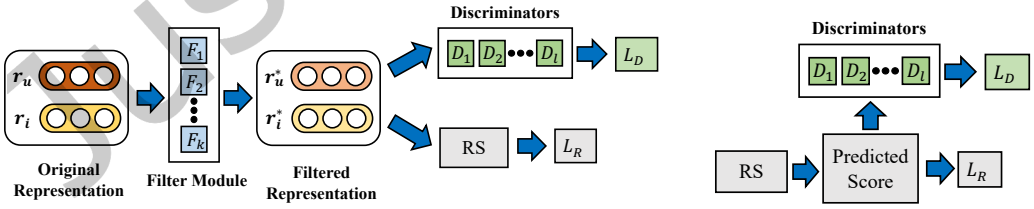
While the above work focuses on the fairness of point-wise predicted scores, Beutal et al. [6] investigate the fairness from the perspective of pair-wise ranking. They demonstrate that the fairness of point-wise ranking tasks does not guarantee the fairness of pair-wise ranking. To improve pair-wise ranking fairness, they add the residual correlations of fairness-related attributes and predicted preferences as regular terms to motivate the model to have similar prediction accuracy across item groups. The loss function is the following Eq.(33) [6]. The second term is the regular fairness term, which will be bigger if the model has a better prediction ability for the clicked item in one group than the other.

$$\begin{aligned} \min_{\theta}(\sum_{(q,j,y,z)\in D} L(f_\theta(q,v_j),(y,z))) + |Corr_P(A,B)| \\ A = (g(f_\theta(q,v_j)) - g(f_\theta(q,v'_j)))(y-y') \\ B = (s_j - s'_j)(y-y') \end{aligned} \tag{33}$$

Here s_j is the binary fairness-related attribute for item j, q is the query consisting of user and context features, y is the user click feedback, z is the post-click engagement, $f_\theta(q,v)$ is the predictions $(\hat{y},\hat{z})$ for item v, $g(\hat{y},\hat{z})$ is the monotonic ranking function from predictions. P is experimental data, and both A and B are random variables over pairs from P.

In summary, we can add a fairness-related regularization term to the loss to improve fairness. Compared to other ranking methods, regularization-based methods are more flexible and easily extensible. However, simply adding regularization terms may make it difficult for the model to learn fairness-related information, which might lead to suboptimal performance.

5.3.2 Adversarial Learning. Several studies use adversarial learning to address the fairness problem [9, 53, 55, 95, 96, 116]. As mentioned earlier, process fairness requires that recommender systems use fair representations. Even though sensitive information is not directly used as input, it may still be indirectly learned by the model into the representation. Adversarial learning is an effective method to reduce the sensitive information in the representation. In addition, it can also be applied to learn fair predicted scores. The basic frameworks of adversarial learning are illustrated in Fig.4.



(a) Learning fair representations (b) Learning fair predicted scores

Fig. 4. Basic frameworks of adversarial learning.

A series of studies [9, 55, 95, 96] are aimed to learn fair representation through adversarial learning. The basic framework is shown in Fig.4(a). Apart from the recommendation model, they often introduce a discriminator

for each fairness-related attribute. These discriminators will predict the corresponding attribute value based on the representations outputted by a filter module which is designed to remove unfair information in original representations. If the discriminator cannot determine the value of these fairness-related attributes according to the filtered representations, the filtered representations will be fair. The learning process can be formalized as the following two-player minimax game.

$$\min_{R} \max_{D} L(R, D) = L_R - \lambda L_D \tag{34}$$

Here L_R is the recommendation loss, and L_D is the attribute prediction loss of discriminators. R is the parameters for the recommendation model, and D is the parameters of discriminators. λ is a hyperparameter. We briefly introduce these methods below.

Avishek and William [9] propose a method to reduce the sensitive information contained in the node representation in graph neural networks, which can be applied to multiple fairness-related attributes simultaneously. This method introduces multiple filter modules in the model, each corresponding to a fairness-related attribute, and removes the corresponding fairness-related attribute information from the node representation. After sensitive information is filtered, all filtered representations of that node are averaged together to obtain a representation without sensitive information v. The discriminator of each fairness-related attribute will predict the corresponding fairness-related attribute of that node based on the representation v. For recommendation, they will only use fair representations v.

In addition to node representations, the network structure around nodes is also important information, which is ignored in the above approach. Wu et al. [96] add discriminators to the graph network recommendation model, which predicts the fairness-related attributes of nodes based on their embeddings and the embeddings of the network structure around the nodes. Experimental results on real datasets also validate that better results can be achieved than the method considering only node information.

The discriminator proposed by Li et al. [55] also predicts the fairness-related attributes of users based on their embeddings. The main difference from previous work is that users can personalize their fairness-related attribute settings.

Unlike the above methods, Wu et al. [95] focus on the fairness of user representations in news recommendations, where the user representation is constructed from the user's reading history. They add a discriminator to the news recommendation model to learn fair user embeddings, which predicts the fairness-related attributes of users based on their embeddings. Besides, they also add an attribute prediction task to learn unfair user embeddings. Furthermore, they add regularization to the loss function to enhance the orthogonality between fair and unfair embeddings.

Apart from learning fair representations, adversarial learning can also be used to learn fair predicted scores. Zhu et al. [116] add discriminator to the recommendation model, which predicts the fairness-related attributes of items based on the predicted scores of the recommendation model. Then they ensure item fairness through the adversary between the recommendation model and the discriminator. The training process can be formalized as the following Eq.(35) [116].

$$\min_{\Theta} \max_{\Psi} \sum_{u \in U} \sum_{i \in I_u^+, j \in I \setminus I_u^+} (L_{BPR}(u, i, j) + \alpha(L_{Adv}(i) + L_{Adv}(j))) + \beta L_{KL} \tag{35}$$

Here $L_{Adv}(i)$ is the log-likelihood loss for an MLP adversary to classify items, and L_{KL} is the KL-loss between the score distribution of each user and a standard normal distribution, which will make the score distribution of each user conform to normal distribution. L_{BPR} is the recommendation loss. Θ and Ψ are learnable parameters for the recommendation model and discriminator. I_u^+ is the set of positive items for user u. α and β are hyperparameters.

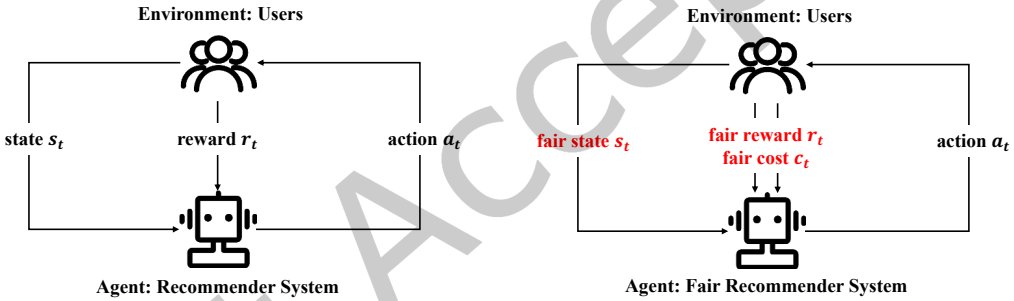
Unlike the above methods that use discriminators to predict fairness-related attributes, Li et al. [53] apply discriminators to reconstruct user/item information. They leverage **textual information** to improve user fairness. The key point is to reduce the mainstream bias in minority representations. The method adds a reconstruction loss to the loss function, requiring the user representation and item representation to restore the original textual information as much as possible. The loss function is as the following Eq.(36) [53].

$$L = L_R + w(L_U + L_I) \tag{36}$$

Here L_R is the rating prediction loss, L_U and L_I are the reconstruction losses for users and items. w is a hyperparameter. They apply convolutional autoencoders to extract and reconstruct textual information.

In summary, current work leverages adversarial learning to learn fair representations or predicted scores to improve recommendation fairness. Adversarial learning is well-suited to fair representation learning and is the dominant approach to this problem. However, since its optimization objective is a minimax optimization problem, it is more difficult to train than the traditional minimization problem.

5.3.3 Reinforcement Learning. Several studies use reinforcement learning (RL) to address the fairness problem [30, 59]. Compared to other methods which mainly consider the immediate fairness impact, reinforcement learning-based fairness methods can optimize fairness in the long run. Fig.5 shows all the differences between existing fair RL methods and general RL methods for the recommendation. Current work mitigates unfairness in recommendations by introducing fairness information in states, rewards, or additional constraints.



(a) RL for recommendation

(b) RL for the fair recommendation

Fig. 5. Illustrations of RL for recommendation and RL for the fair recommendation. Note that (b) shows all the differences between the existing fair RL methods and the general RL methods for the recommendation. In other words, certain fair RL methods may not contain all the differences.

In order to improve long-term fairness, Liu et al. [59] first propose a reinforcement learning-based method. They introduce fairness-related rewards to make recommendations fair. The reward is defined as the following Eq.(37) [59].

$$r_t = \begin{cases} \sum_{i=1}^{l} 1_{A_{c_i}}(a_t)(x_*^i - x_t^i + 1), & if\ y_{a_t} = 1 \\ -\lambda, & if\ y_{a_t} = 0 \end{cases} \tag{37}$$

Here x_*^i is the optimal allocation for group i and x_t^i is the allocation for group i in time step t. A is the item set and A_{c_i} is the item with the attribute value c_i. y_{a_t} is the user feedback on item a_t. λ is a hyperparameter.

They also propose a reinforcement learning-based model based on the actor-critic architecture. The actor-network learns a dynamic fairness-aware ranking strategy vector z, which contains user preferences and the system's fairness status. Then ranking score is calculated based on z and item ID embedding. The critic-network

estimates the value according to z and a fairness allocation vector, which provides information about the current allocation distribution of different groups.

In addition, Ge et al. [30] consider the dynamics in long-term fairness, in other words, the changes of group labels or item attributes due to the user feedback during the whole recommendation process. The dynamic fairness problem is modeled as the Constrained Markov Decision Process, which has been well studied. Specifically, they consider fairness-related constraints to ensure the fairness of recommendations. The constraint is defined as the following Eq.(38).

$$\frac{Exposure_t(G_0)}{Exposure_t(G_1)} \leq \alpha \tag{38}$$

Here $Exposure_t(G_0)$ and $Exposure_t(G_1)$ are the number of exposure in group G_0 and group G_1 at iteration t, and α is a hyperparameter.

They additionally define the cost function as the number of sensitive group items in the recommendation list and find that the fairness constraint can be transformed into a constraint on the cost function. Thus the fairness problem can be formalized as a Markov decision problem with constraints and then solved. They also apply the actor-critic architecture, but the main difference is that their model contains two critics, which approximate the reward and cost, respectively. Compared to the above method containing explicit input about fairness status [59], this model has no fairness-related explicit input.

In summary, existing work on reinforcement learning achieves fair recommendations via modifications to the state, reward, or additional constraints. Compared to other methods, reinforcement learning can optimize long-term and dynamic fairness. Nevertheless, reinforcement learning is difficult to evaluate with offline data and has poor stability [88].

5.3.4 Other Methods. There are also several other fairness ranking methods. Islam et al. [39] use transfer learning to learn fair user representations for career recommendations, and they propose a fair neural model based on neural collaborative filtering (NCF) [37]. They first learn a pre-trained model on insensitive items, then transform the pre-trained user embeddings to mitigate fairness-related attribute biases in them, and then fine-tune them on sensitive items.

Li et al. [56] propose a contextual framework for the fairness-aware recommendation, which is suitable for different fair performance distributions. Specifically, the framework will infer a coefficient for each user/item from the predefined fair distribution. Then the framework will adjust the gradient during the optimization process based on the coefficient.

Borges et al. [8] improve recommendation fairness by adding a stochastic component to a trained VAE model. They find that introducing a normally distributed noise with high variance to the sampling phase can promote fairness despite a slight loss in the recommendation performance.

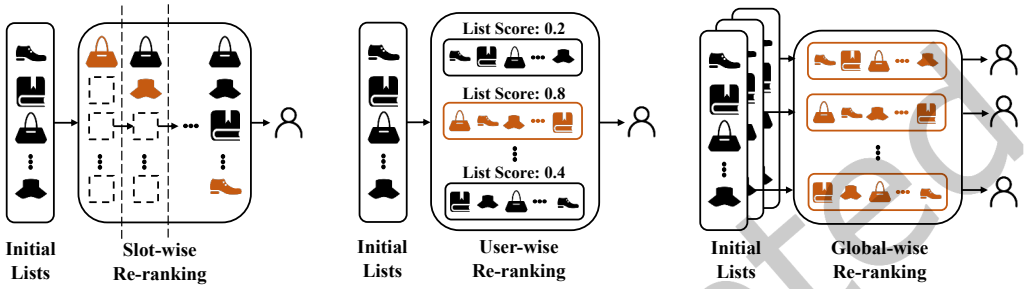
Farnadi et al. [25] propose a rule-based fairness method. They use probabilistic soft logic to implement a fairness-aware hybrid recommender system.

5.4 Re-ranking Methods

Re-ranking methods mainly adjust the outputs of recommendation models to promote fairness. Re-ranking methods have the advantage that their results are nearly identical to the final presentation, making their improvement in outcome fairness the most straightforward. Besides, similar to data-oriented methods, they also have low coupling with the recommender system, as they do not require changing recommendation models. However, because the candidate set in the re-ranking stage is typically small, the performance of re-ranking methods may be hampered. Moreover, they cannot resolve the fair representation issue in the ranking stage.

We divide current re-ranking methods into the following three types: **slot-wise**, **user-wise**, and **global-wise**. Fig.6 illustrates the differences between these three types. The slot-wise re-ranking method re-ranks a

recommendation list by adding items to empty slots in a list one by one. It will select the next item added to the recommendation list by certain rules or re-ranking scores. Unlike slot-wise methods, user-wise re-ranking methods try to directly find the best recommendation list for a user based on the optimization goal of the whole list. While the above two kinds of re-ranking methods are used for a single recommendation list each time, the global-wise re-ranking methods re-rank multiple recommendation lists for multiple users simultaneously. The re-ranking result for one recommendation list may be influenced by other lists.



(a) Slot-wise re-ranking (b) User-wise re-ranking (c) Global-wise re-ranking

Fig. 6. Illustrations for different re-ranking types.

5.4.1 Slot-wise. A few studies [45, 77, 99] propose slot-wise re-ranking methods to improve user fairness in **group recommendations**. Serbos et al. [77] use a greed-based approach to guarantee user fairness in group recommendations. They define the number of satisfied users for a recommendation package P as $SAT_G(P)$. Then the gain from adding a new item i to the current recommendation package P can be defined as $f_G(P,i) = |SAT_G(P \cup i) \backslash SAT_G(P)|$. The recommendation package can be constructed greedily, i.e., start with the empty set and gradually add items to the set to maximize $f_G(P,i)$. Lin et al. [99] also design a greedy algorithm to ensure user fairness in the group recommendation scenario. The difference is that they consider the Pareto efficiency between fairness and recommendation performance. They define the overall recommendation performance as $SW(g,I)$ and the fairness utility as $F(g,I)$. Then the Pareto frontier can be obtained by maximizing $\lambda \times SW(g,I) + (1-\lambda) \times F(g,I)$. Further, recommendations can be obtained by adding items to the recommendation list one by one through a greedy strategy. Similarly, Sacharidis [75] finds Pareto optimal items to promote fairness. After obtaining the candidate Pareto optimal items, they generate ranking scores by linear aggregation strategies and estimate the probability of an item being ranked in Top-K in any strategy. Items are finally ranked based on the estimated probability. While previous studies considered only fixed-length recommendation lists, Kaya et al. [45] consider the fairness of different positions simultaneously. The method greedily selects each item and optimizes the following objective.

$$\begin{aligned} i^* &= \underset{i \in C \backslash OS}{argmax}\, f(i, OS) \\ f(i, OS) &= f(OS \cup \{i\}) - f(OS) \\ f(S) &= \sum_{u \in G} (1 - \prod_{i \in S} (1 - p(rel|u, i))) \end{aligned} \tag{39}$$

Here C is the set of candidate items. OS is the recommendation list recommended to the group. $p(rel|u,i)$ is the probability that item i is relevant to user u.

In the **general recommendation** scenario, existing methods introduce fairness from two main perspectives, one is to maximize utility while satisfying fairness constraints, and another is to optimize both fairness and utility jointly. The former class of methods ensures that the final result is compliant with the fairness constraint, which may entail a relatively large performance loss, while the latter class of methods makes a trade-off between performance and fairness.

Some studies [31, 106] propose algorithms to satisfy the fairness constraint as much as possible at each position. Zehlike et al. [106] propose a priority queue-based approach FA*IR for item fairness scenarios where only two groups exist. FA*IR will maintain two priority queues sorted by relevance, corresponding to two groups. At each position, FA*IR will determine whether the current representation of the protected group satisfies the fairness constraint. If not, select the item with the highest relevance in the protected queue; otherwise, compare the two queues and select the item with the highest relevance. Based on FA*IR, Geyik et al. [31] propose three slot-wise methods to re-rank the results for more than three groups. The first two methods can be considered as extensions of FA*IR. The third algorithm regards fairness constrained re-ranking as an interval constrained sorting problem.

To make a trade-off between fairness and recommendation performance in the re-ranking phase, several studies [44, 58, 83, 84] optimize fairness and utility jointly, and provide hyperparameters to control the loss of recommendation performance. The process of these algorithms can be formalized as the following equation.

$$i^* = \underset{i \in R \setminus C}{argmax}\ \lambda P(u, i) + (1 - \lambda) F(u, i, C) \tag{40}$$

Here R is the set of item candidates. C is the recommendation list for user u, which is empty initially. $P(u, i)$ is the predicted preference of user u to item i. $F(u, i, C)$ is the fairness score. λ is a hyperparameter to control the trade-off between fairness and utility. For each time, these algorithms will select an item i^* from all the available candidate items and then put it into the recommendation list C.

Different studies have different definitions of fairness score $F(u, i, C)$, which is the main difference between them. Steck [84] defines the fairness score from the user perspective and argues that the recommendation should be calibrated for the interests of users, which are measured through interaction history. The fairness score is defined as the KL-divergence between the distribution over different item groups in the history of user u and the distribution over item groups in the recommendation list $C \cup \{i\}$. From the item perspective, Karako and Manggala [44] draw on the ideas of the Maximal Marginal Relevance (MMR) re-ranking algorithm, and they define the fairness score based on item embeddings, which measures how the new item i contributes to the embedding difference between two groups. In addition, considering different diversity tolerance of users, Liu et al. [58] propose a personalized re-ranking method for item fairness. The method of Liu et al. [58] is only available for a single attribute. Based on it, Sonboli et al. [83] find that user tolerance is different across item attributes. They define the personalized fairness score based on multiple item attributes and achieve a better trade-off between fairness and utility.

In the **dynamic ranking** scenario, Morik et al. [67] propose a re-ranking method based on proportional controller. This method also use a linear strategy to combine recommendation performance and fairness. And they theoretically prove that fairness can be guaranteed when the number of rankings is large enough.

$$\sigma_\tau = \underset{d \in D}{argmax}(\hat{R}(d|x) + \lambda err_\tau(d)) \tag{41}$$

Here $\hat{R}(d|x)$ is the estimate relevance for item d to user x and $err_\tau(d)$ is the error term measuring how fairness will be violated if d is recommended.

Further, in the dynamic ranking scenario, Yang and Ai [103] take the marginal fairness into account, i.e., the gain in fairness each time a new item is selected to be added to the recommendation list. They find that the group that maximizes marginal fairness has the lowest current utility-merit ratio. Based on this finding, they propose a probabilistic re-ranking method that jointly optimizes utility and fairness. Specifically, the method will

recommend the most relevant item $\tilde{d}_\tau^k$ in a probability of λ and the fairness-aware item $\overline{d}_\tau^k$ in a probability of $(1-\lambda)$.

$$d_\tau^k \sim (\lambda \tilde{d}_\tau^k + (1-\lambda)\overline{d}_\tau^k) \tag{42}$$

Here d_τ^k is the item selected for k^{th} position of the presented list at time step t, and λ is a hyperparameter.

In summary, the slot-wise methods re-rank independently for each user and add items to the re-ranked list one by one. Compared to other re-ranking methods, slot-wise re-ranking tends to be more intuitive and efficient but shortsighted, which may lead to suboptimal performance.

5.4.2 User-wise. Apart from picking items slot by slot, we can also directly find the recommendation list for a user based on the optimization goal of the whole list. A popular paradigm is **integer programming**. The basic idea is that we can treat some decisions of the re-ranking as decision variables and impose some constraints so that the re-ranking problem can be transformed into an integer programming problem.

In the **group recommendation** scenario, Lin et al. [99] propose an integer programming-based algorithm to ensure user fairness. The integer programming problem can be formalized as the following Eq.43 [99]. The binary decision variables X_i means whether the recommendation set contains the item i. The optimization objective consists of a linear combination of the overall recommendation performance $SW(g, I)$ and the fairness utility $F(g, I)$.

$$\begin{aligned} max \quad & \lambda \times SW(g,I) + (1-\lambda) \times F(g,I) \\ s.t. \quad & \sum_i X_i = K, X_i \in \{0,1\} \end{aligned} \tag{43}$$

Here K is the length of the recommendation list, and λ is a hyperparameter. This problem is NP-hard. They relax X_i to fractional numbers between zero and one to turn it into a convex optimization problem and then select items with the greatest values of X_i.

In **general recommendation** scenario, Biega et al. [7] also formalize the fairness problem as an Integer Linear Programming (ILP) problem. The binary decision variables $X_{i,j}$ means whether the i-th item is placed at the j-th position, and the optimization objective is an amortized unfairness metric calculated through the previous ranking results. Unlike the previous work, they prevent large losses in recommended performance by adding constraints related to recommended performance. The ILP problem is formalized as the following Eq.(44) [7] and then solved by Gurobi, an efficient heuristic algorithm.

$$\begin{aligned} & minimize \sum_{i=1}^{n}\sum_{j=1}^{n} |A_i^{l-1} + w_j - (R_i^{l-1} + r_i^l)| \times X_{i,j} \\ & s.t. \sum_{j=1}^{k}\sum_{i=1}^{n} \frac{2^{r_i^l - 1}}{log_2(j+1)} X_{i,j} \geq \theta \cdot IDCG@k \\ & X_{i,j} \in \{0,1\}, \sum_i X_{i,j} = 1, \sum_j X_{i,j} = 1 \end{aligned} \tag{44}$$

where A_i^{l-1} denotes the cumulative attention value of the ith item in the previous $l-1$ ranking results, and w_j denotes the attention value assigned to the jth position. R_i^{l-1} denotes the cumulative relevance value, and r_i^l denotes the relevance of the ith item in the current ranking. θ is the threshold, which means the changed NDCG is required not to fall below a certain value.

Different from the previous work, Singh and Joachims [80] formalize the problem as a linear programming problem and solve it from a probabilistic ranking perspective. The problem is formalized as the following Eq.(45)

[80]. After the linear programming problem is solved, the final ranking can be sampled through Birkhoff-von Neumann decomposition.

$$\begin{aligned} P &= \arg\max_{P} u^T P v \\ s.t. \mathbf{1}^T P &= \mathbf{1}^T, P\mathbf{1} = \mathbf{1}, 0 \leq P_{i,j} \leq 1, P \text{ is fair} \end{aligned} \tag{45}$$

Here the decision variable $P_{i,j}$ is fractional, which denotes the probability of the item i being placed in the position j. The optimization objective $u^T P v$ is the expected recommendation performance, u_i is the predicted relevance score for item i and v_j is the position coefficient for position j.

In addition to programming-based methods, Mehrotra et al. [64] propose several fairness-aware recommendation strategies. The traditional recommendation strategy will maximize the relevance. Assuming that S is the set of candidate recommendation lists, the traditional recommendation strategy can be formalized as $s_u^* = argmax_{s \in S} \phi(u, s)$, where $\phi(u, s)$ is the relevance estimate function, while a recommendation strategy that considers only fairness is $s_u^* = argmax_{s \in S_u} \psi(s)$, where $\psi(s)$ is the fairness estimate function. To combine fairness with relevance, they propose an interpolation strategy, $s_u^* = argmax_{s \in S_u} (1 - \beta)\phi(u, s) + \beta\psi(s)$, and a probabilistic strategy as the following Eq.(46) [64].

$$s_u^* = \begin{cases} argmax_{s \in S_u} \psi(s) & if p < \beta \\ argmax_{s \in S} \phi(u, s) & otherwise \end{cases} \tag{46}$$

Here β is a hyperparameter.

In summary, the user-wise methods also re-rank independently for each user, and they try to find the optimal list based on the optimization goal of the whole list. Integer programming based on heuristic algorithms is the mainstream method. Compared to slot-wise methods, it considers the information of the whole list to get better performance but is more time-consuming. Besides, compared with the global-wise methods, user-wise methods independently re-rank for each user, which sometimes results in suboptimal performance.

5.4.3 Global-wise. Unlike slot-wise and user-wise methods that re-rank a single recommendation list each time, global-wise methods consider global effects and re-rank multiple lists each time, which are more suitable for solving user fairness problems than the user-wise methods.

Mathematical programming is still a common paradigm. Unlike user-wise methods, the decision variable in global-wise re-ranking methods is usually a binary variable indicating whether an item is recommended to a user. We introduce some representative methods below.

Similar to the user-wise approach, Li et al. [54] propose an integer programming-based approach to solve the user unfairness problem in the general recommendation scenario, which formalizes the problem as the following Eq.(47) [54]. The programming problem is solved by Gurobi.

$$\begin{aligned} &\max_{W_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{N} W_{ij} S_{i,j} \\ &s.t. UGF(Z_1, Z_2, W) < \epsilon, \sum_{j=1}^{N} W_{ij} = K, W_{ij} \in \{0, 1\} \end{aligned} \tag{47}$$

Here the decision variable W_{ij} is the binary variable indicating whether item j is recommended to the user i. $S_{i,j}$ is the preference of user i to item j. Z_1 and Z_2 are two groups of users. UGF is the measurement for user unfairness so that the user group fairness can be guaranteed. K is the length of the recommendation list. ϵ is a hyperparameter.

While previous work focuses on the fairness of the recommendation performance across different users, Fu et al. [28] use integer programming to solve fairness problems in the knowledge-based explainable recommendation. The integer programming problem is similar to Li et al. [54], and they add a fairness constraint to the optimization problem, which controls the unfairness of explanation diversity in the knowledge graphs.

The above studies focus on user fairness. For item fairness, Sürer et al. [87] also propose an integer programming-based method. They first formalize the fairness problem as a 0-1 integer programming problem with provider fairness constraints, then relax the conditions using the Lagrangian method, and finally optimize the problem using the subgradient method.

In addition to programming-based methods, there are also some other re-ranking methods. Mansoury et al. [60, 61] propose a post-processing method for item fairness based on maximum flow matching. The algorithm will build a bipartite graph where the weight between user u and item i is calculated based on the preference of u to i and the degree of i, and then iteratively solve the maximum flow matching problem on the graph. Finally, recommendation lists will be constructed based on the candidates identified by the algorithm. Besides, Zhu et al. [115] propose a parametric post-processing framework for solving the item fairness problem in cold-start scenarios. The method applies an auto-encoder to transform the predicted user preference vector. The transformation needs to satisfy two requirements: the predicted score distribution of under-served items should be as close to the distribution of best-served items as possible, and the predicted score for every user should conform to the same distribution. They propose a generative method and a score scaling method to achieve these requirements.

The above work only considers one-sided fairness. In order to improve **joint fairness**, Patro et al. [71] propose a re-ranking method, which consists of two phases. The first phase greedily assigns the most relevant feasible item to each user's recommendation list with limited exposure to each item in the round-robin manner, which ensures that the exposure of each item is greater than a certain value. The second phase does not limit the exposure to the item and recommends the most relevant item for users who have not received enough recommendations. They theoretically prove that this method can guarantee both envy-free fairness for users and maximin-shared fairness for items.

While the method of Patro et al. [71] guarantees item fairness and user fairness at the individual level, Wu et al. [97] propose an offline re-ranking method and an online method that improve item fairness at the group level and user fairness at the individual level. We introduce the algorithm for the offline version here, as the online version is similar. The algorithm will recommend items for all users from position 1 to position k, i.e., the algorithm will not recommend for a certain position until the positions before it has been recommended. The users are sorted by current recommendation quality (random for the first position). Then the algorithm greedily assigns the most relevant feasible item to each user's recommendation list with limited exposure to each provider. If there is no available item, the position will be skipped. After items in position k are selected, the skipped positions will be recommended with an item with the lowest provider exposure to reduce unfairness further. Experiments show that this algorithm can achieve better fairness than the above algorithm of Patro et al. [71].

In summary, global-wise methods take global effects into account and re-rank multiple lists each time. Since it re-ranks different users simultaneously, it is more suitable for user fairness than other re-ranking methods and tends to achieve better performance in amortized fairness. However, the dependency between different lists makes the re-ranking process difficult to parallelize and more time-consuming.

6 DATASETS FOR FAIRNESS RECOMMENDATION STUDY

6.1 Overview of Fairness Recommendation Datasets

As mentioned in Section 2, most work is aimed at improving group fairness in the recommendation. Group fairness requires certain criteria to divide groups, usually the attributes contained in the dataset, such as the gender of

users. However, not all recommendation datasets have such attribute information, and existing researchers have not paid the same attention to different attributes. For researchers to easily find fairness-related attributes and the relevant datasets, we survey the recommendation datasets used in the previous fairness studies and list the attributes that researchers have considered in their studies. The reviewed datasets are summarized in Table 9.

It is worth mentioning that fairness can also be studied on datasets without attributes. If researchers want to study fairness on attribute-free datasets, there are two options to our knowledge. One is to research fairness issues not requiring additional attributes to divide groups, such as the Rawlsian maximin fairness at the individual level. The other is to manually construct attributes based on interaction information, such as item popularity and user activity. These attribute-free datasets used for fairness studies generally only need to contain user and item ID information and ID-aligned user feedback(e.g., rating, click, and purchase) and may not be limited to the datasets summarized below.

The existing datasets for fairness recommendation studies are relatively rich. As seen in Table 9, there are a relatively large number of recommendation datasets containing attribute information. The scenarios of these datasets are diverse, containing movie recommendations (e.g., *Movielens*, *Flixter*, and *Netflix*), e-commerce recommendations (e.g., *Amazon*, *ModCloth*), and job recommendations (e.g., *Xing*). These datasets contain both large-scale datasets (e.g., *Amazon*) and small-scale datasets (e.g., *Movielens 100K*). The types of interactions in existing datasets are diverse, containing impressions, clicks, and ratings. Moreover, some datasets contain multi-modal information, such as *Amazon* and *Yelp*.

As there is different available information in different scenarios, the attributes considered by researchers are often dataset-specific and vary significantly from one dataset to another, especially for item attributes. For user attributes, gender and age are frequently considered since these attributes are demanded to be fairly treated by anti-discrimination laws. In contrast, item attributes researchers are concerned about are more diverse and contain categories, publishing years, providers, etc.

Apart from these data-specific attributes, there are also some generic attributes to divide groups, such as user activity and item popularity, which only depend on interactions and can be obtained in all datasets. Researchers who cannot use sensitive attributes for some reasons (e.g., privacy) could consider using interaction information to construct these generic attributes. However, it should be noted that such generic attributes are often dynamic, i.e., an individual may belong to different groups at different times. For example, a current popular item may be cold in the previous time, which means it belonged to the protected group previously but is in the unprotected group now [30].

While the existing datasets for fairness studies are diverse, some scenarios and attributes are still worth exploring. For one thing, fairness research can be conducted on some emerging scenarios, such as the short video recommendation scenario, which contains multiple modalities such as video and text. For another, the existing dataset lacks information on some attributes receiving considerable attention, such as race, which is emphasized in anti-discrimination laws [38]. New data may need to be collected to facilitate relevant research, but privacy concerns must also be considered.

Since most work is attribute-based, we present the datasets with fairness-related attributes and the datasets without fairness-related attributes in Sections 6.2 and 6.3, respectively.

6.2 Datasets with Fairness-related Attributes

Amazon. This dataset contains product reviews of various categories from Amazon with user and item profiles, including 142.8 million reviews. For user fairness, previous studies divided user groups based on gender [90] or user activity [28, 54]. Gender information is not directly accessible, so some researchers use the interaction with Clothing products to infer gender identities [90]. The active and inactive users can be grouped based on their number of interactions, total consumption capacity, or maximum consumption level [54]. For item fairness,

Table 9. A lookup table for datasets used in existing fairness research in recommendation. We only list the attributes considered in the previous fairness work as fairness-related attributes in the table, and there may be other attributes in the dataset. "-" represents empty. Datasets are arranged in dictionary order.

Datasets	Fairness-related User Attributes	Fairness-related Item Attributes	Users	Items	Interact.	Related Work
with fairness-related attributes:						
Amazon	activity*, gender	categories, gender of model	20.9M	5.9M	143.6M	[19, 28, 53, 54, 97, 116] [23, 90]
Ctrip Flight	-	airline	3.8K	6K	25.1K	[97]
Flixter	-	popularity*	1M	49K	8.2M	[43]
Google Local	-	business	4.5M	3.1M	11.4M	[71, 97]
Insurance	-	gender, marital status, occupation	1.2K	21	5.3K	[55]
Last.FM 1K	gender, age	-	992	177K	904.6K	[22]
Last.FM 360K	gender, age	-	359.3K	160.1K	17.5M	[22, 61, 71, 96]
ModCloth	body shape	product size	44.7K	1K	99.8K	[90]
Movielens 100K	-	popularity*, provider, year of movie	1K	1.7K	100K	[30, 52, 59, 87, 114]
Movielens 1M	gender, age, occupation	genres, popularity*	6K	3.7K	1M	[30, 39, 55, 60, 96, 115] [12, 25, 45, 73, 75, 116] [22, 43, 61, 99, 104]
Movielens 20M	-	product company, genres	138K	27K	20M	[8, 67, 84, 86, 115]
Sushi	gender, age	seafood or not	5K	100	50K	[43]
Xing	premium/standard	membership, education degree, working country	1.4M	1.3M	8.1M	[19, 56, 115]
Yelp	-	food genres	2.1M	160.5K	8.6M	[62, 77, 116]
without fairness-related attributes:						
BeerAdvocate	-	-	3.7K	37.5K	393K	[53]
CiteULike	-	-	5.5K	16.9K	204.9K	[115]
Epinions	-	-	16.5K	129.3K	512.7K	[60, 76]
KGRec-music	-	-	5.1K	8.6K	751.5K	[45]
Million Song	-	-	1.2M	380K	48M	[8]
Netflix	-	-	480.1K	17.7K	100M	[8]

* User activity and item popularity are not attributes in common sense, but researchers also use them to divide groups as attributes. Thus we add them to the table.

previous studies usually divided item groups according to their categories [116]. A few studies use the gender of the model appearing in product images as a grouping method, which is detected through industrial face detection API [90].

Ctrip Flight. This dataset contains ticket orders on an international flight route from Ctrip with basic information on customers and some ticket information. The entire dataset includes 3.8K customers, 6K kinds of

air tickets and 25K orders. Some researchers treat the airline that the ticket belongs to as the provider, dividing item groups by providers [97].

Flixter. This dataset is a classical movie recommendation dataset and contains 9.1 million movie ratings from Flixter. Some researchers use the item popularity to divide item groups [43]. The movies are first sorted by the interaction number in descending order, then the protected and unprotected groups can be divided according to whether the movie is in the top 1% of the sorted list.

Google Local. This dataset is a location recommendation dataset and contains 11.4 million reviews about 3.1 million local businesses from Google Maps. Some researchers divide item groups based on the business of the reviewed item [71, 97].

Insurance. This dataset is an insurance recommendation dataset in Kaggle, which contains users' information such as gender and occupation. Some researchers divide user groups according to their gender, marital status, and occupation [55].

Last.FM 1K. This dataset is a music recommendation dataset containing 1K play records of 992 users from Last.FM. It contains user demographic information such as gender and age, which are used to divide user groups by some researchers [22].

Last.FM 360K. This dataset is similar to Last.FM 1K but has a larger size, including 17 million play records of 360K users. It also contains the gender and age of users, and some researchers divide user groups based on these attributes [22, 96].

ModCloth. This dataset is an e-commerce recommendation dataset where many products include two human models with different body shapes. The entire dataset contains 100K reviews about 1K clothing products from 44K users. Additionally, there are records of the product sizes which users purchase. For users, some researchers divide users into different body shape groups according to the average size of their purchase [90]. For items, they divide items into different groups according to the body shape of their models [90].

Movielens 100K. This dataset is a classical movie recommendation dataset containing 100K movie ratings with user and item profiles. Some researchers divide items into two groups by item popularity, i.e., the number of exposures for each item [30]. Besides, some studies also divide items into old movies and new movies according to the year of the movie [114], and some researchers randomly assign movies among some providers [59, 87].

Movielens 1M. This dataset is similar to Movielens 100K and has a larger size, including 1 million ratings from 6K users on 4K movies. For users, previous studies divide user groups by their gender, age, and occupation [39, 55, 96]. For items, the movie genres are seen as a fairness-related attribute [116], and item popularity is also considered [30].

Movielens 20M. This dataset is also collected from Movielens, containing 20 million ratings from 138K users on 27K movies. Some studies consider genres [84] and production companies of movies [67] as fairness-related attributes.

Sushi. This dataset includes 5K responses to a questionnaire survey of preference in sushi which contains preference data and demographic data. Some researchers consider three types of fairness-related attributes: age, gender, and whether or not a type of sushi is seafood [43].

Xing. This dataset is a user-view-job dataset, which contains 320M of interactions with user and item profiles such as career level. Some researchers [19] consider the membership, education degree, and working country as fairness-related attributes for items. In addition, whether the user is premium or not is also regarded as a fairness-related attribute for users [56].

Yelp. This dataset is a business review dataset. Some studies only focus on the restaurant business and divide item groups based on the food genres of restaurants [116].

6.3 Datasets without Fairness-related Attributes

We also survey the recommendation datasets without fairness-related attributes in current fairness studies, which are usually used in research on individual fairness.

BeerAdvocate. This dataset [53] contains 1.5 million beer reviews from the BeerAdvocate, including products, user information, and their ratings. Some researchers [53] leverage the reviews in this dataset to enhance the representation of non-mainstream users by adding a textual information reconstruction task.

CiteULike. This dataset [115] includes about 200K records of user preferences toward scientific articles from 5K users. Some researchers [115] utilize this dataset to explore Rawlsian maximin fairness issues in the cold start scenario.

Epinions. This dataset [60, 76] is collected from a Web review site of products, which contains user bidirectional connections and ratings. The whole dataset contains 512K ratings from 16K users on 129K items. Some researchers [60, 76] use this dataset to explore consistent fairness issues at the individual level.

KGRec-music. This dataset [45] is a music recommendation dataset that contains knowledge graphs. The dataset includes about 750K interactions from about 5K users on 8K songs. Some researchers use this dataset [45] to investigate the individual fairness of users in group recommendations.

Million Song. This dataset [8] contains audio attributes and metadata for a million tracks from contemporary popular music, including 1 million songs with 515K dated tracks. Some researchers use this dataset [8] to explore the individual fairness of items.

Netflix. This dataset [8] is a movie recommendation dataset from Netflix, containing 100 million ratings from 480K users over 17K movies. Some researchers [8] leverage this dataset to study the individual fairness of items.

7 FUTURE DIRECTIONS

Fairness is essential for recommender systems and needs to be further exploited. In this section, we discuss some promising future directions for fairness in the recommendation from the perspectives of definition, evaluation, algorithm design, and explanation.

7.1 Definition

A general denition of fairness. As mentioned earlier, many different definitions of fairness have been applied in recommender systems. These fairness definitions may conflict with each other. For example, calibrated fairness may be damaged when ensuring Rawlsian maximin fairness and vice versa. Therefore, it is important to determine the priority between different definitions of fairness, but to our knowledge, there is no work on it yet. In addition, a general definition of fairness may not exist. The appropriate fairness definition may vary in different scenarios. A consensus in each scenario would be helpful.

7.2 Evaluation

Fair comparison between different fairness methods. No effective benchmarks may result in non-reproducible evaluation and unfair comparison and damage the development of the research community. Existing fairness research suffers from this problem since many different fairness measurements and data-processing strategies exist. Hence, it is necessary to propose a standard experimental setting including but not limited to data preprocessing methods, hyper-parameter tuning strategies, and evaluation metrics.

Dataset for new emerging scenarios. Existing datasets for fair recommendation studies are diverse, but there is a lack of investigation on some emerging recommendation scenarios. For example, short video recommendation plays an important role today, and it contains multiple modal information, which is quite different from traditional recommendation scenarios. However, there is a lack of fairness-related work on short-video recommendation

datasets. Whether there are serious unfairness issues in these emerging scenarios and how to address them deserve to be explored.

7.3 Algorithm Design

A win-win for fairness and accuracy. Existing methods often improve the fairness in recommendation with a loss in recommendation performance, and many papers have revealed such a tradeoff between fairness and performance. In the optimal case, fairness is not always in conflict with recommendation performance; for example, for user fairness, both recommendation performance and fairness are optimal if all users receive the most accurate recommendation list. In practice, some work on classification tasks has also found that improving fairness may improve overall accuracy [50]. For industrial recommender systems, the degradation of recommendation performance may result in an unacceptably large loss of revenue. For successfully applying fairness methods to recommender systems, it is necessary to investigate methods to improve fairness while ensuring recommendation accuracy.

Fairness for both user and item. Many approaches to improving fairness have been proposed, but most focus on only one type of user fairness or item fairness. However, both user and item fairness are essential and should be guaranteed in most recommender systems. Hence, it is worthwhile to propose adequate methods for joint fairness. Note that there may be a natural conflict between user fairness and item fairness[97], which also makes joint fairness a challenging topic.

Joint fairness issues can be addressed using different types of methods. First, current practices about data-oriented methods only consider one-sided fairness, and it is worth exploring how to adapt for joint fairness. Second, the joint fairness problem can be regarded as multi-objective learning. The trade-off between multiple fairness and recommendation accuracy might be improved by drawing on Pareto optimization [57] and "seesaw phenomenon" related work such as [89]. Third, most existing re-ranking methods for joint fairness are non-parametric re-ranking algorithms. It is worth investigating how to design learnable re-ranking algorithms, as they have shown better performance on one-sided fairness [115].

Fairness beyond accuracy. Most user fairness work focuses on one evaluation criterion, i.e., accuracy. However, there are many other measurements beyond accuracy, such as diversity, unexpectedness, and serendipity, which are also closely related to user satisfaction. Research has found that unfairness also exists in these measurements [91]. Therefore, we also need to consider more measurements beyond accuracy when ensuring user fairness.

Causal inference for fairness. Eliminating unfairness at the causal level is considered essential and has received a growing interest in machine learning [49, 98]. Similarly, causal inference in recommendation has attracted increasing attention and has become a popular recommendation debiasing technique [108, 110]. However, as we mentioned in Section 2, fairness and bias are different. Only a little work [55] in recommendation focuses on fairness at the causal level. In our opinion, two problems need to be solved. The first problem is how to construct causal graphs for fair recommendations. Most current work focuses on the models based on only ID information. For these models, we can manually design causal graphs. However, for models using additional features, since there exists some causal relationship between these features as well, how to construct causal graphs becomes a challenging problem. The second problem is how to eliminate the influence of unfair factors based on the already constructed causal graphs, especially in the complex causal graphs mentioned above. There is still much room for using causal inference to achieve fair recommendations.

Fairness with missing data. Existing studies usually assume all fairness-related attributes are available in the dataset. However, there are users or items whose fairness-related attributes are missing in many real-world scenarios. For example, some users may fill in their gender as confidential or even false information. In this case, we cannot identify whether the sensitive group is treated unfairly, and existing fairness methods will be ineffective.

Therefore, it is necessary to investigate methods to improve fairness when fairness-related information is missing. Solving this problem also helps to reduce the risk of sensitive information leakage since we only depend on partial information. There has been some related work on classification tasks [17, 18, 48, 92, 100], while in recommender systems, it is still a problem to be explored.

Fairness in a real system. Industrial recommender systems usually consist of three phases: recall, ranking, and re-ranking. Some objectives other than accuracy, such as diversity, are often involved in the re-ranking phase. Existing studies have found that some post-processing techniques, such as diversity re-ranking, may increase user unfairness [84], implying that some fairness methods applied before re-ranking will be ineffective. Therefore, it is necessary to investigate how to improve fairness more effectively in real-world systems. For example, we can consider adding fairness-oriented recall in the recall phase. On the other hand, industrial recommender systems usually require a short response time, so more efficient re-ranking algorithms need to be proposed. Moreover, industrial recommender systems often have multiple objectives, such as reading time and purchase. However, the fairness of the corresponding model for each goal cannot guarantee the fairness of the final recommendations [93], which also poses a challenge for applying fairness in industrial systems.

7.4 Explanation

What are the causes of unfairness? Explainability is crucial for recommender systems as it can improve the persuasiveness of recommendations, increase user satisfaction, and enhance the transparency of the whole system [109]. Explaining why unfairness occurs can deepen the understanding of unfairness and facilitate the design of more effective fairness methods. Although many approaches have been developed to improve fairness in the recommendation, there is relatively little work on the explainability of fairness, i.e., why unfairness occurs. Only a few studies [113] have theoretically proven that a specific class of models leads to unfairness in recommendation results. Causal inference-based and more theoretical analyses remain a challenge.

8 CONCLUSION

Unfairness is widespread in recommender systems, which has attracted increasing attention in recent years, and a series of fairness definitions, measurements, and methods have been proposed. This survey systematically reviews fairness-related research in the recommendation and summarizes current fairness work from multiple perspectives, including definitions, views, measurements, datasets, and methods.

For fairness definitions, previous studies mainly focus on outcome fairness, which we further classify according to different targets and concepts. We find that group fairness is the most common target, and consistent fairness and calibrated fairness are the most common concepts. As for fairness views, we present some views to classify fairness issues in the recommendation, containing fairness subjects, fairness granularity, and fairness optimization objects. For fairness measurements, we introduce representative measurements of existing work and summarize common metrics of different fairness definitions. As for fairness methods, we review representative studies from data-oriented methods, ranking methods, and re-ranking methods. It is common for researchers to develop ranking and re-ranking methods to achieve fair recommendations, while there are only a few studies on adjusting data to improve fairness. Additionally, we summarize fairness-related recommendation datasets occurring in previous fairness work for researchers to find relevant datasets easily.

Furthermore, we discuss some promising future research directions from different perspectives for fairness in the recommendation. In terms of definitions, which definition of fairness is most proper to recommender systems may be an important problem for future work. As for evaluation, we could develop effective benchmarks to compare different fairness methods fairly. In terms of algorithm design, we discuss some promising future work containing fairness methods for both user and item and fairness methods beyond accuracy. In terms of

explanation, explaining why unfairness exists could be a problem worth exploring. Finally, we hope this survey may help readers better understand fairness issues in recommender systems and provide some inspiration.

REFERENCES

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

[2] J Stacy Adams. 1963. Towards an understanding of inequity. *The journal of abnormal and social psychology* 67, 5 (1963), 422.

[3] W. D. Ross Aristotle and Lesley Brown. 2009. *The Nicomachean ethics*. Oxford University Press.

[4] Warren Ashby. 1950. Teleology and Deontology in Ethics. *The Journal of Philosophy* 47, 26 (1950), 765–773. http://www.jstor.org/stable/2020659

[5] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[6] 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 *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2212–2220.

[7] 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* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 405–414. https://doi.org/10.1145/3209978.3210063

[8] 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* (Limassol, Cyprus) *(MEDES '19)*. Association for Computing Machinery, New York, NY, USA, 95–102. https://doi.org/10.1145/3297662.3365798

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

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

[11] Robin Burke, Jackson Kontny, and Nasim Sonboli. 2018. Synthetic attribute data for evaluating consumer-side fairness. *arXiv preprint arXiv:1809.04199* (2018).

[12] 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.

[13] Carlos Castillo. 2019. Fairness and transparency in ranking. In *ACM SIGIR Forum*, Vol. 52. ACM New York, NY, USA, 64–71.

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

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

[16] 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).

[17] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 339–348. https://doi.org/10.1145/3287560.3287594

[18] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair Transfer Learning with Missing Protected Attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 91–98. https://doi.org/10.1145/3306618.3314236

[19] Yashar Deldjoo, Vito Walter Anelli, Hamed Zamani, Alejandro Bellogín, and Tommaso Di Noia. 2019. Recommender systems fairness evaluation via generalized cross entropy. *arXiv preprint arXiv:1908.06708* (2019).

[20] Fernando Diaz, Bhaskar Mitra, Michael D. Ekstrand, Asia J. Biega, and Ben Carterette. 2020. *Evaluating Stochastic Rankings with Expected Exposure*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3340531.3411962

[21] 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* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[22] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 172–186. http://proceedings.mlr.press/v81/ekstrand18b.html

[23] Michael D. Ekstrand, Mucun Tian, Mohammed R. Imran Kazi, Hoda Mehrpouyan, and Daniel Kluver. 2018. Exploring Author Gender in Book Rating and Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 242–250. https://doi.org/10.1145/3240323.3240373

[24] Boli Fang, Miao Jiang, Pei-yi Cheng, Jerry Shen, and Yi Fang. 2020. Achieving Outcome Fairness in Machine Learning Models for Social Decision Problems.. In *IJCAI*. 444–450.

[25] 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).

[26] Bruce Ferwerda, Mark Graus, Andreu Vall, Marko Tkalcic, and Markus Schedl. 2016. The influence of users' personality traits on satisfaction and attractiveness of diversified recommendation lists. In *EMPIRE 2016 Emotions and Personality in Personalized Systems: Proceedings of the 4th Workshop on Emotions and Personality in Personalized Systems co-located with ACM Conference on Recommender Systems (RecSys 2016)*, Vol. 1680. CEUR-WS, 43–47.

[27] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)* 14, 3 (1996), 330–347.

[28] 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.

[29] David García-Soriano and Francesco Bonchi. 2021. Maxmin-Fair Ranking: Individual Fairness under Group-Fairness Constraints. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 436–446. https://doi.org/10.1145/3447548.3467349

[30] 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. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 445–453. https://doi.org/10.1145/3437963.3441824

[31] 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.

[32] Mohammad Ghodsi, Mohammadtaghi Hajiaghayi, Masoud Seddighin, Saeed Seddighin, and Hadi Yami. 2018. Fair Allocation of Indivisible Goods: Improvements and Generalizations. In *Proceedings of the 2018 ACM Conference on Economics and Computation* (Ithaca, NY, USA) *(EC '18)*. Association for Computing Machinery, New York, NY, USA, 539–556. https://doi.org/10.1145/3219166.3219238

[33] 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* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1808–1812. https://doi.org/10.1145/3404835.3463235

[34] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 903–912. https://doi.org/10.1145/3178876.3186138

[35] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence* (New Orleans, Louisiana, USA) *(AAAI'18/IAAI'18/EAAI'18)*. AAAI Press, Article 7, 10 pages.

[36] Ananya Gupta, Eric Johnson, Justin Payan, Aditya Kumar Roy, Ari Kobren, Swetasudha Panda, Jean-Baptiste Tristan, and Michael Wick. 2021. Online Post-Processing in Rankings for Fair Utility Maximization. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 454–462. https://doi.org/10.1145/3437963.3441724

[37] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, Rick Barrett, Rick Cummings, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 173–182. https://doi.org/10.1145/3038912.3052569

[38] Elisa Holmes. 2005. Anti-Discrimination Rights Without Equality. *Modern Law Review* 68 (02 2005). https://doi.org/10.1111/j.1468-2230.2005.00534.x

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

[40] Jacob Jacoby. 1984. Perspectives on Information Overload. *Journal of Consumer Research* 10, 4 (1984), 432–435. http://www.jstor.org/stable/2488912

[41] Rajendra K Jain, Dah-Ming W Chiu, William R Hawe, et al. 1984. A quantitative measure of fairness and discrimination. *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA* (1984).

[42] Benjamin Johnson and Richard Jordan. 2017. Why Should Like Cases Be Decided Alike? A Formal Model of Aristotelian Justice.

[43] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2018. Recommendation independence. In *Conference on Fairness, Accountability and Transparency*. PMLR, 187–201.

[44] Chen Karako and Putra Manggala. 2018. Using image fairness representations in diversity-based re-ranking for recommendations. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*. 23–28.

[45] 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* (Virtual Event, Brazil) *(RecSys '20)*. Association for Computing Machinery, New York, NY, USA, 101–110. https://doi.org/10.1145/3383313.3412232

[46] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.

[47] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 100–109.

[48] Ömer Kırnap, Fernando Diaz, Asia Biega, Michael Ekstrand, Ben Carterette, and Emine Yilmaz. 2021. Estimation of Fair Ranking Metrics with Incomplete Judgments. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 1065–1075. https://doi.org/10.1145/3442381.3450080

[49] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4069–4079.

[50] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. In *34th Conference on Neural Information Processing Systems*. Curran Associates, Inc.

[51] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 182 (Nov. 2019), 26 pages. https://doi.org/10.1145/3359284

[52] Jurek Leonhardt, Avishek Anand, and Megha Khosla. 2018. User fairness in recommender systems. In *Companion Proceedings of the The Web Conference 2018*. 101–102.

[53] Roger Zhe Li, Julián Urbano, and Alan Hanjalic. 2021. Leave No User Behind: Towards Improving the Utility of Recommender Systems for Non-Mainstream Users. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 103–111. https://doi.org/10.1145/3437963.3441769

[54] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-Oriented Fairness in Recommendation. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 624–632. https://doi.org/10.1145/3442381.3449866

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

[56] Yangkun Li, Mohamed-Laid Hedia, Weizhi Ma, Hongyu Lu, Min Zhang, Yiqun Liu, and Shaoping Ma. 2022. Contextualized Fairness for Recommender Systems in Premium Scenarios. *Big Data Research* 27 (2022), 100300. https://doi.org/10.1016/j.bdr.2021.100300

[57] Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qingfu Zhang, and Sam Kwong. 2019. *Pareto Multi-Task Learning*. Curran Associates Inc., Red Hook, NY, USA.

[58] Weiwen Liu, Jun Guo, Nasim Sonboli, Robin Burke, and Shengyu Zhang. 2019. Personalized Fairness-Aware Re-Ranking for Microlending. In *Proceedings of the 13th ACM Conference on Recommender Systems* (Copenhagen, Denmark) *(RecSys '19)*. Association for Computing Machinery, New York, NY, USA, 467–471. https://doi.org/10.1145/3298689.3347016

[59] Weiwen Liu, Feng Liu, Ruiming Tang, Ben Liao, Guangyong Chen, and Pheng Ann Heng. 2020. Balancing between accuracy and fairness for interactive recommendation with reinforcement learning. *Advances in Knowledge Discovery and Data Mining* 12084 (2020), 155.

[60] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. FairMatch: A Graph-Based Approach for Improving Aggregate Diversity in Recommender Systems. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 154–162. https://doi.org/10.1145/3340631.3394860

[61] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2021. A Graph-Based Approach for Mitigating Multi-Sided Exposure Bias in Recommender Systems. *ACM Trans. Inf. Syst.* 40, 2, Article 32 (nov 2021), 31 pages. https://doi.org/10.1145/3470948

[62] Masoud Mansoury, Bamshad Mobasher, Robin Burke, and Mykola Pechenizkiy. 2019. Bias disparity in collaborative recommendation: Algorithmic evaluation and comparison. *arXiv preprint arXiv:1908.00831* (2019).

[63] 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.

[64] 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* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 2243–2251. https://doi.org/10.1145/3269206.3272027

[65] Martin Mladenov, Elliot Creager, Omer Ben-Porat, Kevin Swersky, Richard Zemel, and Craig Boutilier. 2020. Optimizing long-term social welfare in recommender systems: A constrained matching approach. In *International Conference on Machine Learning*. PMLR, 6987–6998.

[66] Natwar Modani, Deepali Jain, Ujjawal Soni, Gaurav Kumar Gupta, and Palak Agarwal. 2017. Fairness aware recommendations on behance. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 144–155.

[67] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. *Controlling Fairness and Bias in Dynamic Learning-to-Rank*. Association for Computing Machinery, New York, NY, USA, 429–438. https://doi.org/10.1145/3397271.3401100

[68] 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.

[69] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, Vol. 1170.

[70] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proceedings of the 23rd International Conference on World Wide Web* (Seoul, Korea) *(WWW '14)*. Association for Computing Machinery, New York, NY, USA, 677–686. https://doi.org/10.1145/2566486.2568012

[71] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P. Gummadi, and Abhijnan Chakraborty. 2020. FairRec: Two-Sided Fairness for Personalized Recommendations in Two-Sided Platforms. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1194–1204. https://doi.org/10.1145/3366423.3380196

[72] Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys* 3 (2009), 96–146.

[73] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. 2019. Fighting Fire with Fire: Using Antidote Data to Improve Polarization and Fairness of Recommender Systems. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 231–239. https://doi.org/10.1145/3289600.3291002

[74] John Rawls. 2020. *A theory of justice*. Harvard university press.

[75] Dimitris Sacharidis. 2019. Top-N Group Recommendations with Fairness. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing* (Limassol, Cyprus) *(SAC '19)*. Association for Computing Machinery, New York, NY, USA, 1663–1670. https://doi.org/10.1145/3297280.3297442

[76] Dimitris Sacharidis, Carine Pierrette Mukamakuza, and Hannes Werthner. 2020. Fairness and Diversity in Social-Based Recommender Systems. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (Genoa, Italy) *(UMAP '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 83–88. https://doi.org/10.1145/3386392.3397603

[77] Dimitris Serbos, Shuyao Qi, Nikos Mamoulis, Evaggelia Pitoura, and Panayiotis Tsaparas. 2017. Fairness in Package-to-Group Recommendations. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 371–379. https://doi.org/10.1145/3038912.3052612

[78] 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*. Association for Computing Machinery, New York, NY, USA, 4094–4103. https://doi.org/10.1145/3459637.3481948

[79] Hyejin Shin, Sungwook Kim, Junbum Shin, and Xiaokui Xiao. 2018. Privacy Enhanced Matrix Factorization for Recommendation with Local Differential Privacy. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1770–1782. https://doi.org/10.1109/TKDE.2018.2805356

[80] 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* (London, United Kingdom) *(KDD '18)*. Association for Computing Machinery, New York, NY, USA, 2219–2228. https://doi.org/10.1145/3219819.3220088

[81] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 5427–5437. https://proceedings.neurips.cc/paper/2019/hash/9e82757e9a1c12cb710ad680db11f6f1-Abstract.html

[82] Nasim Sonboli and Robin Burke. 2019. Localized Fairness in Recommender Systems. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP'19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 295–300. https://doi.org/10.1145/3314183.3323845

[83] Nasim Sonboli, Farzad Eskandanian, Robin Burke, Weiwen Liu, and Bamshad Mobasher. 2020. Opportunistic Multi-aspect Fairness through Personalized Re-ranking. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 239–247.

[84] Harald Steck. 2018. Calibrated Recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 154–162. https://doi.org/10.1145/3240323.3240372

[85] Ana-Andreea Stoica, Christopher Riederer, and Augustin Chaintreau. 2018. Algorithmic Glass Ceiling in Social Networks: The Effects of Social Recommendations on Network Diversity. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 923–932. https://doi.org/10.1145/3178876.3186140

[86] Maria Stratigi, Jyrki Nummenmaa, Evaggelia Pitoura, and Kostas Stefanidis. 2020. Fair Sequential Group Recommendations. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing* (Brno, Czech Republic) *(SAC '20)*. Association for Computing Machinery, New York, NY, USA, 1443–1452. https://doi.org/10.1145/3341105.3375766

[87] Özge Sürer, Robin Burke, and Edward C. Malthouse. 2018. Multistakeholder Recommendation with Provider Constraints. In *Proceedings of the 12th ACM Conference on Recommender Systems* (Vancouver, British Columbia, Canada) *(RecSys '18)*. Association for Computing Machinery, New York, NY, USA, 54–62. https://doi.org/10.1145/3240323.3240350

[88] Kanata Suzuki and Tetsuya Ogata. 2020. Stable Deep Reinforcement Learning Method by Predicting Uncertainty in Rewards as a Subtask. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 23–27, 2020, Proceedings, Part II* (Bangkok, Thailand). Springer-Verlag, Berlin, Heidelberg, 651–662. https://doi.org/10.1007/978-3-030-63833-7_55

[89] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. *Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations*. Association for Computing Machinery, New York, NY, USA, 269–278. https://doi.org/10.1145/3383313.3412236

[90] 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* (Houston, TX, USA) *(WSDM '20)*. Association for Computing Machinery, New York, NY, USA, 618–626. https://doi.org/10.1145/3336191.3371855

[91] Ningxia Wang and Li Chen. 2021. *User Bias in Beyond-Accuracy Measurement of Recommendation Algorithms*. Association for Computing Machinery, New York, NY, USA, 133–142. https://doi.org/10.1145/3460231.3474244

[92] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael I Jordan. 2020. Robust optimization for fairness with noisy protected groups. *arXiv preprint arXiv:2002.09343* (2020).

[93] 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* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 436–444. https://doi.org/10.1145/3437963.3441732

[94] Leonard Weydemann, Dimitris Sacharidis, and Hannes Werthner. 2019. Defining and Measuring Fairness in Location Recommendations. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Recommendations, Geosocial Networks and Geoadvertising* (Chicago, Illinois) *(LocalRec '19)*. Association for Computing Machinery, New York, NY, USA, Article 6, 8 pages. https://doi.org/10.1145/3356994.3365497

[95] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairness-aware News Recommendation with Decomposed Adversarial Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4462–4469.

[96] 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* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 2198–2208. https://doi.org/10.1145/3442381.3450015

[97] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. 2021. TFROM: A Two-sided Fairness-Aware Recommendation Model for Both Customers and Providers. *arXiv preprint arXiv:2104.09024* (2021).

[98] Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual Fairness: Unidentification, Bound and Algorithm. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (Macao, China) *(IJCAI'19)*. AAAI Press, 1438–1444.

[99] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-Aware Group Recommendation with Pareto-Efficiency. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) *(RecSys '17)*. Association for Computing Machinery, New York, NY, USA, 107–115. https://doi.org/10.1145/3109859.3109887

[100] Shen Yan, Hsien-te Kao, and Emilio Ferrara. 2020. *Fair Class Balancing: Enhancing Model Fairness without Observing Sensitive Attributes*. Association for Computing Machinery, New York, NY, USA, 1715–1724. https://doi.org/10.1145/3340531.3411980

[101] Forest Yang, Mouhamadou Cisse, and Oluwasanmi O Koyejo. 2020. Fairness with overlapping groups; a probabilistic perspective. *Advances in Neural Information Processing Systems* 33 (2020).

[102] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management* (Chicago, IL, USA) *(SSDBM '17)*. Association for Computing Machinery, New York, NY, USA, Article 22, 6 pages. https://doi.org/10.1145/3085504.3085526

[103] Tao Yang and Qingyao Ai. 2021. Maximizing Marginal Fairness for Dynamic Learning to Rank. In *Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. Association for Computing Machinery, New York, NY, USA, 137–145. https://doi.org/10.1145/

3442381.3449901

[104] Sirui Yao and Bert Huang. 2017. Beyond parity: Fairness objectives for collaborative filtering. *arXiv preprint arXiv:1705.08804* (2017).

[105] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web* (Perth, Australia) *(WWW '17)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1171–1180. https://doi.org/10.1145/3038912.3052660

[106] 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.

[107] Meike Zehlike and Carlos Castillo. 2020. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 2849–2855. https://doi.org/10.1145/3366424.3380048

[108] Wenhao Zhang, Wentian Bao, Xiao-Yang Liu, Keping Yang, Quan Lin, Hong Wen, and Ramin Ramezani. 2020. *Large-Scale Causal Approaches to Debiasing Post-Click Conversion Rate Estimation with Multi-Task Learning*. Association for Computing Machinery, New York, NY, USA, 2775–2781. https://doi.org/10.1145/3366423.3380037

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

[110] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. *Causal Intervention for Leveraging Popularity Bias in Recommendation*. Association for Computing Machinery, New York, NY, USA, 11–20. https://doi.org/10.1145/3404835.3462875

[111] Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. 2021. Contrastive Learning for Debiased Candidate Generation in Large-Scale Recommender Systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 3985–3995. https://doi.org/10.1145/3447548.3467102

[112] Qiliang Zhu, Qibo Sun, Zengxiang Li, and Shangguang Wang. 2020. FARM: A Fairness-Aware Recommendation Method for High Visibility and Low Visibility Mobile APPs. *IEEE Access* 8 (2020), 122747–122756. https://doi.org/10.1109/ACCESS.2020.3007617

[113] Ziwei Zhu, Yun He, Xing Zhao, Yin Zhang, Jianling Wang, and James Caverlee. 2021. Popularity-Opportunity Bias in Collaborative Filtering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining* (Virtual Event, Israel) *(WSDM '21)*. Association for Computing Machinery, New York, NY, USA, 85–93. https://doi.org/10.1145/3437963.3441820

[114] 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* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 1153–1162. https://doi.org/10.1145/3269206.3271795

[115] 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* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 767–776. https://doi.org/10.1145/3404835.3462948

[116] Ziwei Zhu, Jianling Wang, and James Caverlee. 2020. *Measuring and Mitigating Item Under-Recommendation Bias in Personalized Ranking Systems*. Association for Computing Machinery, New York, NY, USA, 449–458. https://doi.org/10.1145/3397271.3401177