A Survey of Research on Fair Recommender Systems

Article *in* Journal of Information Processing and Management · May 2022

Yashar Deldjoo
Politecnico di Bari

A Survey of Research on Fair Recommender Systems

Yashar Deldjoo[a], Dietmar Jannach[b], Alejandro Bellogin[c], Alessandro Diffonzo[a], Dario Zanzonelli[a]

[a] *Polytechnic University of Bari, Italy*
[b] *University of Klagenfurt, Austria*
[c] *University Autonomous of Madrid, Spain*

Abstract

Recommender systems can strongly influence which information we see online, e.g, on social media, and thus impact our beliefs, decisions, and actions. At the same time, these systems can create substantial business value for different stakeholders. Given the growing potential impact of such AI-based systems on individuals, organizations, and society, questions of *fairness* have gained increased attention in recent years. However, research on fairness in recommender systems is still a developing area. In this survey, we first review the fundamental concepts and notions of fairness that were put forward in the area in the recent past. Afterwards, we provide a survey of how research in this area is currently operationalized, for example, in terms of the general research methodology, fairness metrics, and algorithmic approaches. Overall, our analysis of recent works points to certain research gaps. In particular, we find that in many research works in computer science very abstract problem operationalizations are prevalent, which circumvent the fundamental and important question of what represents a fair recommendation in the context of a given application.

Keywords: Recommender Systems, Fairness

1. Introduction

Recommender systems (RS) are one of the most visible and successful applications of AI technology in practice, and personalized recommendations—as provided on many modern e-commerce or media sites–can have a substantial impact on different stakeholders. On e-commerce sites, for example, the choices of consumers can be largely influenced by recommendations, and these choices are often directly related to the profitability of the platform. On news websites or social media, on the other hand, personalized recommendations may determine to a large extent which information we see, which in turn may shape not only our own beliefs, decisions, and actions, but also the beliefs of a community of users or an entire society.

In academia, recommenders have historically been considered as "benevolent" systems that create value for consumers, e.g., by helping them find relevant items, and that this value for consumers then translates to value for businesses, e.g., due to higher sales numbers or increased customer retention [1]. Only in the most recent years, more

Email address: deldjooy@acm.org (Yashar Deldjoo)

awareness was raised regarding possible negative effects of automated recommendations, e.g., that they may promote items on an e-commerce site that mainly maximize the profit of providers or that they may lead to an increased spread of misinformation on social media.

Given the potentially significant effects of recommendations on different stakeholders, researchers increasingly argue that providing recommendations may raise various ethical questions and should thus be done in a *responsible* way [2]. One important ethical question in this context is that of the *fairness* of a recommender system, see [3, 4], reflecting related discussions on the more general level of *fair machine learning* and *fair AI* [5, 6].

During the last years, researchers have discussed and analyzed different dimensions in which a recommender system should be fair, or vice versa, may lead to a lack of fairness. Given the nature of fairness as a social construct, it however seems difficult (or even impossible [4]) to establish a general definition of what represents a fair recommendation. Beside the subjective nature of fairness, there are also often competing interests of different stakeholders to be considered in real-world recommendation settings [7].

With this survey, our goal is to provide an overview of what has been achieved in this emerging area so far and to highlight potential research gaps. Specifically, drawing on an analysis of more than 130 recent papers in computer science, we investigate: *(i)* which dimensions and definitions of fairness in RS have been identified and established, *(ii)* at which application scenarios researchers target and which examples they provide, and *(iii)* how they operationalize the research problem in terms of methodology, algorithms, and metrics. Based on this analysis, we then paint a landscape of current research in various dimensions and discuss potential shortcomings and future directions for research in this area.

Overall, we find that research in computing typically assumes that a clear definition of fairness is available, thus rendering the problem as one of designing algorithms to optimize a given metric. Such an approach may however appear too abstract and simplistic, cf. [8], calling for more faceted and multi-disciplinary approaches to research in fairness-aware recommendation.

2. Background: Fairness in Recommender Systems

2.1. Examples of Unfair Recommendations

In the general literature in Fair ML/AI, a key use case is the automated prediction if a convicted criminal will recidivate. In this case, an ML-based system is usually considered unfair if its predictions depend on demographic aspects like ethnicity and when it then ultimately discriminates members of certain ethnic groups. In the context of our present work, such use cases of ML-based decision-support systems are not in the focus. Instead, we focus on common application areas of RS where *personalized* item suggestions are made to users, e.g., in e-commerce, media streaming, or news and social media sites.

At first sight, one might think that the recommendation providers here are independent businesses and it is entirely at their discretion which shopping items, movies, jobs, or social connections they recommend on their platforms. Also, one might assume that the *harm* that is made by such recommendations is limited, compared, e.g., to the legal decision problem mentioned above. There are, however, a number of situations also in common application scenarios of RS where many people might think that a system is unfair in some sense. For example, an e-commerce platform might be considered unfair if it mainly promotes those shopping items that maximize its own profit but not consumer utility. Besides such intentional interventions, there might also be situations where an RS reinforces existing discrimination patterns or biases in the data, e.g., when a system on an employment platform mainly recommends lower-paid jobs to certain demographic groups.

Questions of fairness in RS are however not only limited to the consumer's side. In reality, a recommendation service often involves multiple stakeholders [7]. On a music streaming platforms, for example, we not only have the consumers, but also the artists, record labels and the platform itself, which might have diverging goals that may be affected by the recommendation service. Artists and labels are usually interested to increase their visibility through the recommendations. Platform providers, on the other hand might seek to maximize the engagement with the service across the entire user base, which might result in promoting mostly already popular artists and tracks with the recommendations. Such a strategy however easily leads to a "rich-get-richer" effect and reduces the chances of less popular artists to be exposed to consumers, which might be considered *unfair to providers*. Finally, there are also use cases where recommendations may have *societal* impacts, in particular on news and social media sites. Some may for example consider it unfair if a recommender system only promotes content that emphasizes one side of a political discussion or promotes misinformation that is suitable to discriminate certain user groups.

Some of the discussed examples of unfair recommendations might appear to be rather *ethical* or *moral* questions or related to an organization's business model, e.g., when an e-commerce provider does not optimize for consumer value or when niche artists are not frequently exposed to users. However, note that being fair in the bespoke examples may also serve providers, e.g., when consumers establish long-term trust due to valuable recommendations or when they engage more with a music service when they discover more niche content. Finally, there are also *legal* guardrails that may come into play, e.g., when a large platform uses a monopoly-like market position to put certain providers into an inappropriately bad position. The current draft of the European Commission's Digital Service Act[1] can be seen as a prime example where recommender systems and their potential harms are explicitly addressed in legal regulations.

Overall, a number of examples exist where recommendations might be considered unfair for different stakeholders. In the context of the survey presented in this work, we are particularly interested in which specific *real-world* problems related to unfair recommendations are considered in the existing literature.

2.2. Reasons for Unfairness

There are different reasons why a recommender system might exhibit a behavior that may be considered unfair, see [4] and [9]. One common issue mentioned in the literature is that the data on which the machine learning model

[1] https://eur-lex.europa.eu/legal-content/en/TXT/?uri=COM:2020:825:FIN

is trained is biased. Such biases might for example be the result of the specifics of the data collection process, e.g., when a biased sampling strategy is applied. A machine learning model may then "pick up" such a bias and reflect it in the resulting recommendations.

Another source of unfairness may lie in the machine learning model itself, e.g., when it even reinforces existing biases or existing skewed distributions in the underlying data. Differences between recommendation algorithms in terms of reinforcing popularity biases and concentration effects were for example examined in [10]. In some cases, the machine learning model might also directly consider a "protected characteristic" (or a proxy thereof) in its predictions [4]. To avoid discrimination, and thus unfair treatment, of certain groups, a machine learning model should therefore not make use of protected characteristics such as age, color, or religion.

Unfairness that is induced by the underlying data or algorithms may arise unknowingly to the recommendation provider. It is however also possible that a certain level of unfairness is designed into a recommendation algorithm, e.g., when a recommendation provider aims to maximize monetary business metrics while at the same time keeping users satisfied as much as possible [11, 12]. Likewise, a recommendation provider may have a political agenda and particularly promote the distribution of information that mainly supports their own viewpoints.

Some works finally mention that the "world itself may be unfair or unjust" [4], e.g., due to historical discrimination of certain groups. In the context of *algorithmic* fairness—which is the topic of our present work—such historical developments are however often not in the focus. Rather, the question is to what extent this is reflected into the data or how this unfairness influences the fairness goals, e.g., by implementing *affirmative action* policies. where the goal is to support traditionally underrepresented groups.

In general, the underlying reasons also determine *where* in a machine learning pipeline[2] interventions can or should be made to ensure fairness (or to mitigate unfairness). In a common categorization [5, 13, 14, 15], this could be achieved *(i)* in a data pre-processing phase, *(ii)* during model learning and optimization, and *(iii)* in a post-processing phase. In particular, in the model learning and post-processing phase, fairness-ensuring algorithmic interventions must be guided by an *operationalizable* (i.e., mathematically expressed) goal. In case of affirmative action policies, one could for example aim to have an equal distribution of recommendations of members of the majority group and members of an underrepresented group. As we will see in Section 4, such a goal is often formalized as a target distribution and/or in the form of an evaluation metric to gauge the level of existing or mitigated fairness.

2.3. Notions of Fairness

When we deal with phenomena of unfairness like those described, and when our goal is to avoid or mitigate such phenomena, the question naturally arises what we consider to be fair, in general and in a specific application context. Fairness, in general, fundamentally is a societal construct or a *human value*, which has been discussed for centuries in

[2]In [9], Ashokan and Hass review where biases may occur in a typical machine learning pipeline from data generation, over model building and evaluation, to deployment and user interaction.

many disciplines like philosophy and moral ethics, sociology, law, or economics. Correspondingly, countless definitions of fairness were proposed in different contexts, see for example Verma et al. [16, 17] for a high-level discussion of the definition of fairness in machine learning and ranking algorithms, or Mulligan et al. [18] for the relationship to social science conception of fairness.

One popular characterization can be found in [5], where fairness in the context of decision making is considered as the "*absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits*". This definition captures two common notions of fairness that are used in the recommender systems literature, where often a differentiation between *individual* fairness and *group* fairness is made. Individual fairness roughly expresses that similar individuals should be treated similarly, e.g., candidates with similar qualifications should be ranked similarly in a job recommendation scenario. How to determine similarity is key here for fairness and protected characteristics like religion or gender should not be factors that make candidates dissimilar. Group fairness, in contrast, aims to ensure that "similar groups have similar experience" [4]. Typical groups in such a context are a majority or dominant group and a protected group (e.g., an ethnical minority). Questions of group fairness were traditionally discussed in fair ML research in the context of classification problems, and are often referred to as different forms of *statistical parity*. A fair classifier would therefore assign a member of a protected and an unprotected class with equal probability to the "positive" class, e.g., the class that is assumed to pay back a loan.

An in-depth discussion of these—sometimes even incompatible—notions of fairness is beyond the scope of this work, which focuses on an analysis of how scholars in recommender systems operationalize the research problem. For questions of individual fairness, this might relate to the problem of defining a similarity function. For certain group fairness goals, on the other hand, one has to determine which are the (protected) attributes that determine group membership. Furthermore, it is often required to define/indicate precisely some *target distributions*. Later, in Section 4, where we review the current literature, we will introduce additional notions of fairness and their operationalizations as they are found in the studied papers. As we will see, a key point here is that researchers often propose to use very abstract operationalizations (e.g., in the form of fairness metrics), which was identified earlier as a potential key problem in the broader area of fair ML in [8].

2.4. Related Concepts: Responsible Recommendation and Biases

Issues of fairness are often discussed within the broader area of *responsible* recommendation [19, 20]. In [19], the authors in particular discuss potential negative effects of recommendations and their underlying reasons with a focus on the media domain. Specific phenomena in this domain include the emergence of filter bubbles and echo chambers. There are, however, also other more general potential harms such as popularity biases as well as fairness-related aspects like discrimination that can emerge in media recommendation settings. Fairness is therefore seen as a particular aspect of responsible recommendation in [19]. A similar view is taken in [20], where the authors review a number of related concerns of responsibility: accountability, transparency, safety, privacy, and ethics. In the context of our present work, most of these concepts are however only of secondary interest.

More important, however, is the use of the term *bias* in the related literature. As discussed above, one frequently discussed topic in the area of recommender systems is the problem of *biased data* [21, 22]. One issue in this context is that the data that is collected from existing websites—e.g., regarding which content visitors view or what consumers purchase—may not be "natural" but biased by what is shown to users through an already existing recommender system. This, in turn, then may lead to biased recommendations when machine learning models reflect or reinforce the bias, as mentioned above. In works that address this problem, the term bias is often used in a more statistical sense as done in [20]. However, the use of the term is not consistent in the literature, as observed also in [21] and in our work. In some early papers, bias is used almost synonymously with fairness. In [23], for example, bias is used to "*refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others*". In our work, we acknowledge that biased recommendations may be unfair, but we do not generally equate bias with unfairness. Considering the problem of popularity bias in recommender systems, such a bias may lead to an over-proportional exposure of certain items to users. This, however, not necessarily leads to unfairness in an ethical or legal sense.

3. Research Methodology

In this section, we first describe our methodology of identifying relevant papers for our survey. Afterwards, briefly discuss how our survey extends previous works in this area.

3.1. Paper Collection Process

We identified relevant research papers in a systematic way [24] by querying digital libraries with predefined search terms. Based on our prior knowledge about the literature, we used the following search terms in order to cover a wide range of works in an emerging area, where terminology is not yet entirely unified: *fair recommend*, *fair collaborative system*, *fair collaborative filtering*, *bias recommend*, *debias recommend*, *fair ranking*, *bias ranking*, *unbias ranking*, *re-ranking recommend*, *reranking recommend*. To identify papers, we queried DBLP and the ACM Digital Library[3] in their respective search syntax, stating that the provided keywords must appear in the title of the paper.

From the returned results, we then removed all papers that were published as preprints on arXiv.org.[4] and we removed survey papers. We then manually scanned the remaining 234 papers. In order to be included in this survey, a paper had to fulfill the following additional criteria:

- It had to be explicitly about *fairness*, at least by mentioning this concept somewhere in the paper. Papers which, for example, focus on mitigating popularity biases, but which do not mention that fairness is an underlying goal of their work, were thus not considered.

[3] https://dl.acm.org, https://dblp.org

[4] Note that DBLP indexes arXiv papers.

- It had to be about *recommender systems*. Given the inclusiveness of our set of query terms, a number of papers were returned which focused on fair information retrieval. Such works were also excluded from our study.

This process left us with 130 papers. The papers were read by at least two researchers and categorized in various dimensions, see Section 4.

3.2. Relation to Previous Surveys

A number of related surveys were published in the last few years. The recent monograph by Ekstrand et al. [4] discusses fairness aspects in the broader context of *information access* systems, an area which covers both information retrieval and recommender systems. Their comprehensive work in particular includes a taxonomy of various fairness dimensions, which also serves as a foundation of our present work. The survey provided in [21] focuses on biases in recommender systems, and connects different types of biases, e.g., popularity biases, with questions of fairness, see also [25]. A categorization of different types of biases is provided in the work along with a review of existing approaches to bias mitigation. Both works, [4] and [21], are different from our present work as our goal is not to provide a novel categorization of fairness concepts or algorithms used in the literature. Instead, our main goal is to investigate the current state of existing research, e.g., in terms of which concepts and algorithmic approaches are predominantly investigated and where there might be research gaps.

Different survey papers were published also in the more general area of fair machine learning or fair AI, as mentioned above [5, 6]. Clearly, many questions and principles of fair AI apply also to recommender systems, which can be seen as a highly successful area of applied machine learning. Differently from such more general works, however, our present work focuses on the particularities of fairness in recommender systems.

4. A Landscape of Research

In this section, we categorize the identified literature along different dimensions to paint a landscape of current research and to identify existing research gaps.

4.1. Publication Activity per Year

Interest in fairness in recommender systems has been constantly growing through the past few years. Figure 1 shows the number of paper per year that were considered in our survey. Questions of fairness in information retrieval have been discussed for many years, see, e.g., [26] for an earlier work. In the area of recommender systems, however, the earliest paper we identified through our search, which only considers papers in which fairness is *explicitly* addressed, was only published in 2017.

4.2. Types of Contributions

Academic research on recommender systems in general is largely dominated by algorithmic contributions, and we correspondingly observe a large amount of new methods that are published every year. Clearly, building an effective

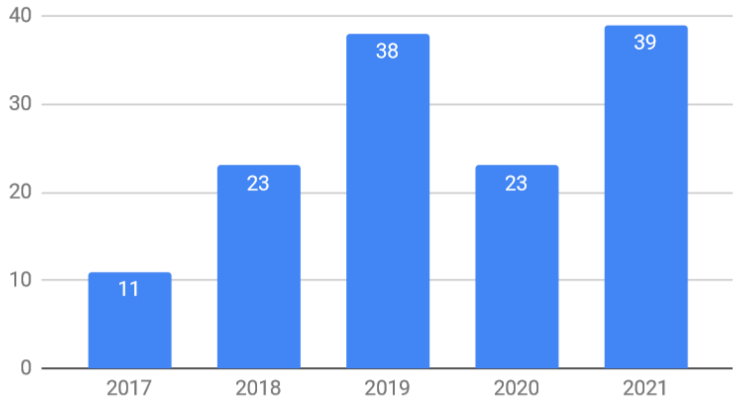

Figure 1: Number of papers published per year.

recommender system requires more than a smart algorithm, e.g., because recommendation to a large extent is also a problem of human-computer interaction and user experience design [27, 28]. Now when questions of fairness should be considered as well, the problem becomes even more complex as for example ethical questions may come into play and we may be interested on the impact of recommendations on individual stakeholders, including society.

In the context of our study, we were therefore interested in which *general types* of contributions we find in the computer science and information systems literature on fair recommendation. Based on the analysis of the relevant papers, we first identified two general types of works: (a) *technical* papers, which, e.g., propose new algorithms, protocols, and metrics or analyze data, and (b) *conceptual* papers. The latter class of papers is diverse and includes, for example, papers that discuss different dimensions of fair recommendations, papers that propose conceptual frameworks, or works that connect fairness with other quality dimensions like diversity.

We then further categorized the technical papers in terms of their *specific technical type* of contribution. The main categories we identified are (a) *algorithm* papers, which for example propose re-ranking techniques, (b) *analytic* papers, which for example study the outcomes of a given algorithm, and (c) *methodology* papers, which propose new metrics or evaluation protocols.

Figure 2 shows how many papers in our survey were considered as technical and conceptual papers. Non-technical papers cover a wide range of contributions, such as guidelines for designers to avoid *compounding* previous injustices [29], exploratory studies that investigate user perceptions of fairness [30], or discussions about how difficult it is to audit these types of systems [31].

We observe that today's research on fairness on recommender systems is dominated by technical papers. In addition, we find that the majority of these works focuses on improved algorithms, e.g., to debias data or to obtain a fairer recommendation outcome through list re-ranking. To some extent this is expected as we focus on the computer science literature. However, we have to keep in mind that the concepts of fairness and unfairness or social constructs may depend on a variety of environmental factors in which a recommender system is deployed. As such, the research

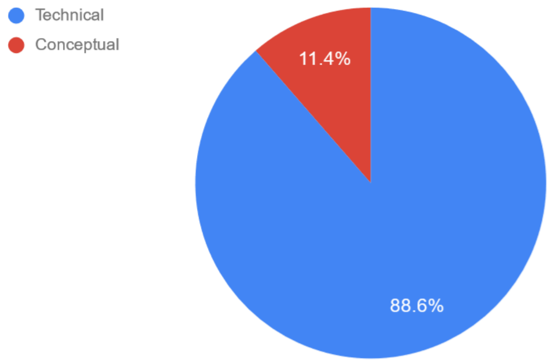

Figure 2: Technical vs. Conceptual Papers.

focus in the area of fair recommender systems seems rather narrow and on algorithmic solutions. As we will observe later, however, such algorithmic solutions commonly assume that a pre-existing and mathematically defined optimization goal is available, e.g., a target distribution of recommendations. In practical applications, the major challenges mostly lie (a) in establishing a common understanding and agreement on such a fairness goal and (b) in finding or designing an operationalizable optimization goal (e.g., a computational metric) which represents a reliable measure or proxy for the given fairness goal.

4.3. Notions of Fairness

In [32], a taxonomy of different notions of fairness was introduced. In the following, we review the literature following this taxonomy.

Group Fairness vs. Individual. A very common differentiation in fair recommendation is to distinguish between *group* fairness and *individual* fairness, as indicated before. With group fairness, the goal is to achieve some sort of *statistical parity* between *protected* groups [33]. In fair machine learning, a traditional goal often is to ensure that there are equal number of members of each protected group in the outcome, e.g., when it comes to make a ranked list of job candidates. The protected groups in such situations are commonly determined by characteristics like age, gender, or ethnicity. Achieving individual fairness in the described scenario means that candidates with similar characteristics should be treated similarly. To operationalize this idea, therefore some distance metric is needed to assess the similarity of individuals. This can be a challenging task, since there is no consensus on the notion of similarity, and it could be task-specific [34]. Ideas of individual fairness in machine learning were discussed in an early work in [34], where it was also observed that achieving group fairness might lead to an unfair treatment at the individual level. In the candidate ranking example, favoring members of protected groups to achieve parity might ultimately result in the non-consideration of a better qualified candidate from a non-protected group. As a result, group and individual fairness are frequently viewed as trade-offs, which is not always immediately evident [33].

Figure 3 shows how many of the surveyed papers focus on each category. The figure shows that research on

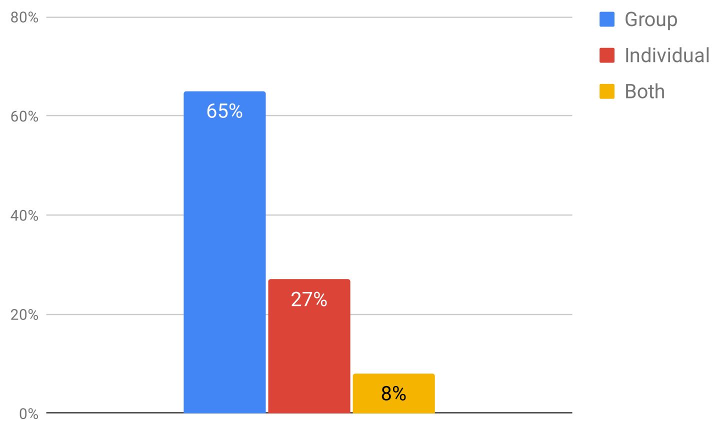

Figure 3: Individual vs. Group Fairness.

scenarios where group fairness is more common than works that adopt the concept of individual fairness. Only in rare cases, both types of fairness are considered.

Group fairness entails comparing, on average, the members of the privileged group against the unprivileged group. One overarching aspect to identify research papers on groups fairness is the distinction between the *(i)* benefit type (exposure vs. relevance), and *(ii)* major stakeholders (consumer vs. provider). Exposure relates to the degree to which items or item groups are exposed uniformly to all users/user groups. Relevance (accuracy) indicates how well an item's exposure is effective, i.e., how well it meets the user's preference. For recommender systems, where users are first-class citizens, there are multiple stakeholders, consumers, producers and side stakeholders (see next section).

To perform fairness evaluation for item recommendation tasks, the users or items are divided into non-overlapping groups (segments) based on some form of *attributes*. These attributes can be either supplied externally by the data provider (e.g., gender, age, race) or computed internally from the interaction data (e.g., based on user activity level, mainstreamness, or item popularity). Nonetheless, we give the most frequently used features in the recommendation fairness literature to operationalize. In Table 1, we provide a list of the most commonly used attributes in the recommendation fairness literature, which can be utilized to operationalize the group fairness concept. They are divided according to Consumer fairness (C-Fairness), Producer Fairness (P-Fairness), and combinations (CP-Fairness).

Moreover, in the area of recommender systems, a number of *people recommendation* scenarios can be identified that are similar to classical fair ML problems. These include recommenders on dating sites, social media sites that provide suggestions for connections, and specific applications, e.g., in the educational context [57]. In these cases, user demographics may play a major role. However, in many other cases, e.g., in e-commerce recommendation or media recommendation, it is not always immediately clear what protected groups may be. In [59] and other works, for example, user groups are defined based on their activity level, and it is observed that highly active users (of an e-commerce site) receive higher-quality recommendations in terms of usual accuracy measures. This is in general not

Table 1: Overview of common attributes used when addressing fairness concepts from the perspectives of consumers, providers, or both.

Goal 1: Consumer Fairness	Attribute
Target: *Demographic parity* – sensitive attributes are attained by birth and not under a user's control.	• Gender [35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49] • Race [50, 40, 51, 52, 49] • Age [35, 53, 54, 55, 38] • Nationality [56]
Target: *Merit-based fairness* – attained through a user's merit over time.	• Education [54, 57] • Income [54]
Target: *Behavior-oriented fairness* – attained based on a user's engagement with the system/item catalog.	• User (in)activeness [58, 59, 60, 61, 62] • User (non)mainstreaminess [25, 63]
Target: *Other emerging attributes*	• Physio/psychological [41, 64] • Sentiment-based [65]
Goal 2: Provider Fairness	
Target: *Item producer/creator* – sensitive attribute based on who the item producer is.	• News author [66], music artist [67], movie director [68]
Target *Producer's demographic or general information* – sensitive attribute based on to which demographic group the item producer belongs, e.g., male vs. female artists.	• Gender [69, 68, 70, 71], geographical region [57]
Target: *Item information* – sensitive attribute based on the item information itself.	• Price and brand [35, 72], geographical region [73, 48]
Target: *Interaction-oriented fairness* – sensitive attribute based on the interactions observed on items e.g., popularity.	• Popularity [35, 74, 75, 76, 77, 78, 79, 56, 80, 81], cold items [82]
Target: *Other emerging attributes*	• Premium membership [83], sentiment and reputation [65, 84]
Target: *Non-sensitive attributes*	• Movie and music genre [43, 45, 85, 67]
Goal 3: Consumer Provider Fairness	
Target: Combinations of two targets from C-Fairness and P-Fairness.	• Same category of sensitive attributes for both users and items (e.g. *behavior-oriented*) [86, 87, 65, 80, 48] • Different categories of sensitive attributes [35, 83, 44, 43, 56, 71]

surprising because there is more information a recommender system can use to make suggestions for more active users. However, it stands to question if an algorithm that returns the best recommendations it can generate given the available amount of information should be considered unfair. Recent studies have also focused on two-sided CP-Fairness, as illustrated in [86, 87]. In these works, the authors demonstrate the existence of inequity in terms of exposure to popular products and the quality of recommendation offered to active users. It is unknown if increasing fairness on one or both sides (consumer/producers) has an effect on the overall quality of the system. In [86], an optimization-based re-ranking strategy is then presented that leverages consumer and provider-side benefits as constraints. The authors demonstrate that it is feasible to boost fairness on both the user and item sides without compromising (and even enhancing) recommendation quality.

Different from traditional fairness problems in ML, research in fairness for recommenders also frequently considers the concept of *fairness towards items* or their suppliers, see also [32], which differentiates between user and item fairness. In these cases, items *can* be related to users, for example artists on a music recommendation scenario, but they can also be arbitrary objects. In these research works, the idea often is to avoid an unequal (or: unfair) *exposure* of items of different groups. In some works, e.g., [88], the *popularity* of items is considered an important attribute, and the goal is to give fair exposure to items that belong to the long tail. In other research works that focus on fair item exposure, e.g., in [89], groups are defined based on attributes that are in practice not protected, e.g., the price range of an accommodation; often, also synthetic data is used. The purpose of such experiments is usually to demonstrate the effectiveness of an algorithm if (any) groups were given. Nonetheless, in these cases it often remains unclear in which ways evaluations make sense with datasets from domains where there is no clear motivation for considering questions of fairness. Also, in cases where the goal is to increase the exposure of long-tail items, no particular motivation is usually provided about why recommending (already) popular items is generally unfair. There are often good reasons why certain items are unpopular and should not be recommended, for example, simply because they are of poor quality [90].

Fairness for items at the *individual* level, in particular for cold-start items, is for example discussed in [82]. In general, as shown in Figure 3, works that consider aspects of individual fairness are rather rare and the definition from classical fair ML settings—similar individuals should be treated similarly—can not always be directly transferred to recommendation scenarios. In [91], for example, the goal is to make sure that the system is not able to derive a user's sensitive attribute, e.g., gender, and should thus be able to treat male and female individuals similarly. Most other works that focus on individual fairness address problems of *group recommendation*, i.e., situations where a recommender is used to make item suggestions for a group of users. Group recommendation problems have been studied for many years [92, 93], usually with the goal to make item suggestions that are acceptable for all group members and where all group members are treated similarly. In the past, these works were often not explicitly mentioning fairness as a goal, because this was an implicit underlying assumption of the problem setting. In more recent works on group recommendation, in contrast, fairness is explicitly mentioned, e.g., in [64, 94, 95], maybe also due to the current interest in this topic. Notable works in this context are [64] and [96], which are one of the few works in our survey which consider questions of fairness *perceptions*.

Finally, we underline the resurgence of the notion of *calibration recommendation* or *calibration fairness* in recommender systems. In ML, calibration is a fundamental concept which occurs when the expected proportions of (predicted) classes match the observed proportions data points in the available data. Similarly, the purpose of calibration fairness is to reflect a measure of the deviation of users' interests from the suggested recommendation in an acceptable proportion [97, 98, 99].

Single-sided and Multi-Sided Fairness. Traditionally, research in computer science on recommender systems has focused on the consumer value (or utility) of recommender systems, e.g., on how algorithmically generated suggestions may help users deal with information overload. Providers of recommendation services are however primarily interested in the value a recommender can ultimately create for their organization. The organizational impact of recommender systems has been, for many years, the focus in the field of information systems, see [100] for a survey. Only in recent years we observe an increased interest on such topics in the computer science literature. Many of these recent works aim to shed light on the impact of recommendations in a multistakeholder environment, where typical stakeholders may include consumers, service providers, suppliers of the recommendable items, or even society [7, 101].

In multistakeholder environments, there may exist trade-offs between the goals of the involved entities. A recommendation that is good for the consumer might for example not be the best for the profit perspective of the provider [12]. In a similar vein, questions of fairness can be viewed from multiple stakeholders, leading to the concept of *multisided* fairness [102]. As mentioned above, there can be fairness questions that are related to the suppliers of the items. Again, there can also be tradeoffs, i.e., what may be a fair recommendation for users might be in some ways be seen to be unfair to item suppliers, e.g., when their items get limited exposure.

Figure 4 shows the distribution of works that focus on one single side of fairness and works which address questions of multisided fairness. The illustration clearly shows that the large majority of the works concentrates on the single-sided case, indicating an important *research gap* in the area of multisided fairness within multistakeholder application scenarios.

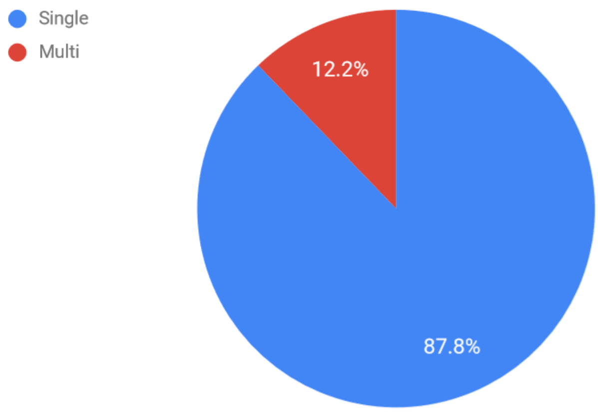

Figure 4: Fairness Notions: Single-sided vs. Multi-sided Fairness.

Among the few studies on multi-sided fairness, [103] discusses techniques for CP-fairness in matching platforms such as Airbnb and Uber. Patro et al. [104] model the fair recommendation problem as a constrained fair allocation problem with indivisible goods and propose a recommendation algorithm that takes producer fairness into consideration. Wu et al. [105] propose an individual-based perspective, where fairness is defined as the same exposure for all producers and the same NDCG for all consumers involved.

Static vs. Dynamic Fairness. Another dimension of fairness research relates to the question whether the fairness assessment is done in a static or dynamic environment [32]. In static settings, the assessment is done at a single point of time, as commonly done also in offline evaluations that focus on accuracy. Thus, it is assumed that the attributes of the items do not change, that the set of available items does not change, and that the analysis that is made at one point in time is sufficient to assess the fairness of algorithms or if an unfairness mitigation technique is effective.

Such static evaluations however have their shortcomings, e.g., as there may be feedback loops that are induced by the recommendations. Also, some effects of unfairness and the effects of corresponding mitigation strategies might only become visible over time. Such longitudinal studies require alternative evaluation methodologies, for example, approaches based on synthetic data or different types of *simulation*, such as those developed in the context of reinforcement learning algorithms, see [106, 107, 11, 108, 109] for simulation studies and related frameworks in recommender systems.

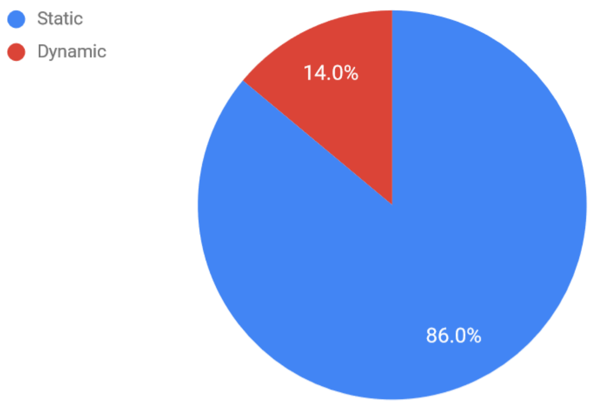

Figure 5: Fairness Notions: Static vs. Dynamic Evaluation.

Figure 5 shows how many studies in our survey considered static and dynamic evaluation settings, respectively. Static evaluations are clearly predominant: we only found 12 works that consider dynamically changing environments. In [78], for example, the authors consider the dynamic nature of the recommendation environment by proposing a fairness-constrained reinforcement learning algorithm so that the model dynamically adjusts its recommendation policy to ensure the fairness requirement is satisfied even when the environment changes. A similar idea is developed in [73], where a long-term balance between fairness and accuracy is considered for interactive recommender systems, by incorporating fairness into the reward function of the reinforcement algorithm. On the other hand, works such as [110] and [35] model fairness in a specific snapshot of the system, by simply taking the system and its training information as a fixed image of the interactions performed by the users on the system.

Associative vs. Causal Fairness. The final categorization discussed in [32] contrasts *associative* and *causal* fairness. One key observation by the authors in that context is that most research in fair ML is based on association-based (correlation-based) approaches. In such approaches, researchers typically investigate the potential *"discrepancy of statistical metrics between individuals or subpopulations*". However, certain aspects of fairness cannot be investigated properly without considering potential causal relations, e.g., between a sensitive (protected) feature like gender and the model's output. In terms of methodology, causal effects are often investigated based on counterfactual reasoning [111, 112].

Figure 6 shows that there are only *three* works investigating recommendation fairness problems based on causality considerations. In [113], for example, the authors propose the use of counterfactual explanation to provide fair recommendations in the financial domain. An interesting alternative is presented in [112], where the authors analyze the causal relations between the protected attributes and the obtained results.

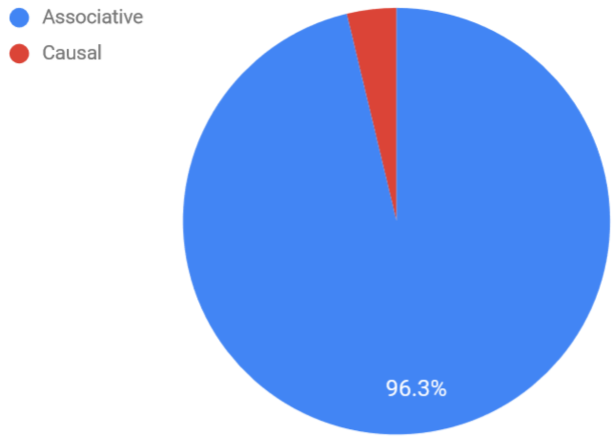

Figure 6: Fairness Notions: Associative vs. Causal Fairness.

One additional dimension we have discovered through our literature analysis is the use of *constrained-based approaches* to integrate or model fairness characteristics in recommender systems. For example, [58] address the issue of enforcing equality to biased data by formulating a constrained multi-objective optimization problem to ensure that sampling from imbalanced sub-groups does not affect gradient-based learning algorithms; the same work and others—including [114] or [115]—define fairness as another constraint to be optimized by the algorithms. In [115], for example, such a constraint is amortized fairness-of-exposure.

4.4. Application Domains and Datasets

Next, we look at application domains that are in the focus of research on fair recommendations. Figure 7 shows an overview of the most frequent application domains and how many papers focused on these domains in their evaluations. The by far most researched domain is the recommendation of videos (movies) and music, followed by e-commerce and finance. For many other domains shown in the figure (e.g., jobs, tourism, or books), only a few papers were identified. Certain domains were only considered in one or two papers. These papers are combined in the "Other" domain in Figure 7.

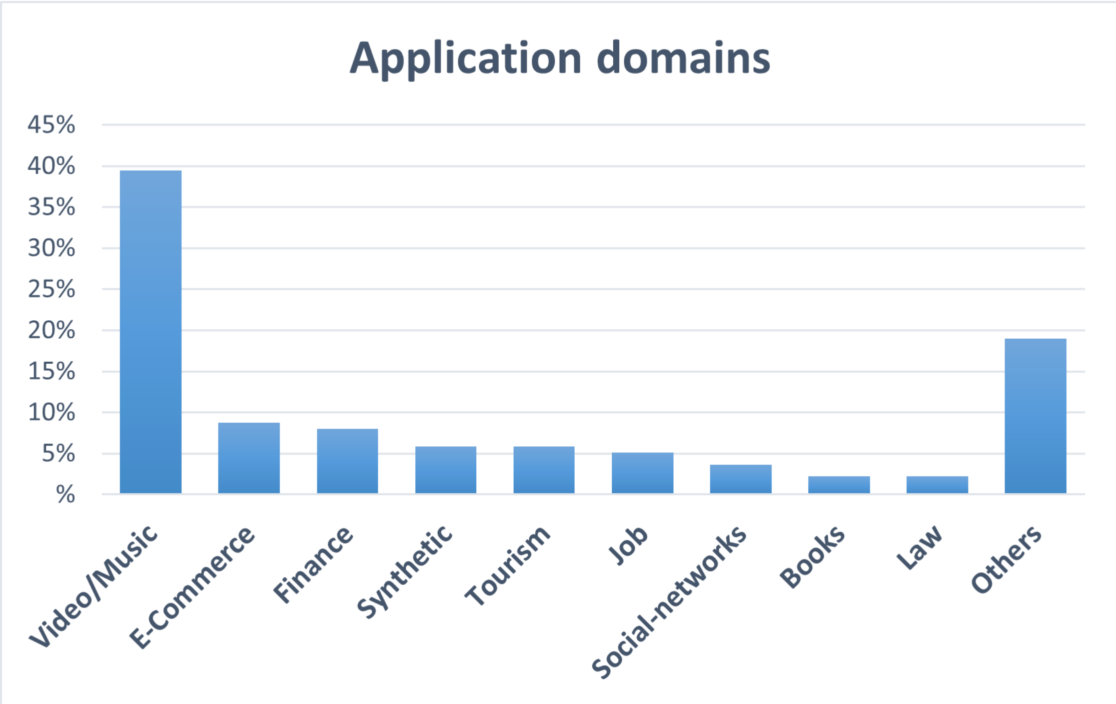

Figure 7: Application Domains.

Since most of the studied papers are technical papers and use an offline experimental procedure, corresponding datasets from the respective domains are used. Strikingly often, in more than one third of the papers, one of the MovieLens datasets is used. This may seem surprising as some of these datasets not even contain information about sensitive attributes. Generally, these observations reflect a common pattern in recommender systems research, which is largely driven by the availability of datasets. The MovieLens datasets are a notorious case and have been used for all sorts of research in the past [116]. Fairness research in recommender systems thus seems to have a quite different focus than fair ML research in general, which is often about avoiding discrimination of people.

We may now wonder which specific fairness problems are studied with the help of the MovieLens rating datasets. What would be unfair recommendations to users? What would be unfair towards the movies (or their providers)? It turns out that item popularity is often the decisive attribute to achieve *fairness towards items*, and quite a number of works aim to increase the exposure of long-tail items which are not too popular, see, e.g., [74]. In terms of *fairness towards users*, the technical proposal in [75] for example aims to serve users with recommendations that reflect their past diversity preferences with respect to movie genres. An approach towards *fairness to groups* is proposed in [117]. Here, groups are not identified by their protected attribute, but by the recommendation accuracy that is achieved (using any metric) for the members of the group.

Continuing our discussions above, such notions of unfairness may not be undisputed. When some users receive recommendations with lower accuracy, this might be caused by their limited activity on the platform or their unwillingness to allow the system to collect data. Actually, one may consider it unfair to artificially lower the quality of recommendations for the group of highly active and open users. Also, a movie may simply not be popular, because it is of poor quality, as mentioned above. It is not clear why recommending to many users would make the system fairer, and equating bias (or skewed distributions) with unfairness in general seems questionable. Finally, also for the user fairness calibration approach from [75] it is less than clear why diversifying recommendations according to user tastes would increase the system's fairness. It may increase the quality of the recommendations, but a system that generates lower-quality recommendations for everyone is probably not one we would call unfair.

In several cases it therefore seems that the addressed problem settings are not too realistic or artificial. One main reason for this phenomenon in our view lies in the lack of suitable datasets for domains where fairness really matters. These could for example be the problem of job recommendations on business networks or people recommendations on social media which can be discriminatory. In today's research, often datasets from rather non-critical domains or synthetic datasets are used to showcase the effectiveness of a technical solution [78, 63, 118, 117, 58, 43, 79, 46, 119]. While this may certainly be meaningful to demonstrate the effects of, e.g., a fairness-aware re-ranking algorithm, such research may appear to remain quite disconnected from real-world problems. This phenomenon of an "abstraction trap" was discussed earlier in [120].

4.5. Methodology

In this section, we review how researchers approach the problems from a methodological perspective.

Research Methods. In principle, research in recommender systems can be done through experimental research (e.g., with a field study or through a simulation) or non-experimental research (e.g., through observational studies or with qualitative methods) [121]. In recommender systems research, three main types of experimental research are common: (a) offline experiments based on historical data, (b) user studies (laboratory studies), and (c) field tests (A/B tests, where different systems versions are evaluated in the real world). Figure 8 shows how many papers fall into each category. Like in general recommender systems research [122], we find that offline experiments are the predominant form of research. Note that we here only consider technical papers, and not the conceptual or theoretical ones that we identified. Only in very few cases, humans were involved in the experiments, and in even fewer cases we found reports of field tests. Regarding user studies, [64] for example involves real users to evaluate fairness in a group recommendation setting. On the other hand, notable examples of field experiment are provided in [46], where a gender-representative re-ranker is deployed for a randomly chosen 50% of the recruiters on the *LinkedIn Recruiter* platform (A/B testing), and in [110]. We only found one paper that relied on interviews as a qualitative research method [30]. Also, only very few papers used more than one experiment type, e.g., [123] were both a user study and an offline experiment were conducted.

The dominance of offline experiments points to a research gap in terms of our understanding of *fairness perceptions* by users. Many technical papers that use offline experiments assume that there is some target distribution or a

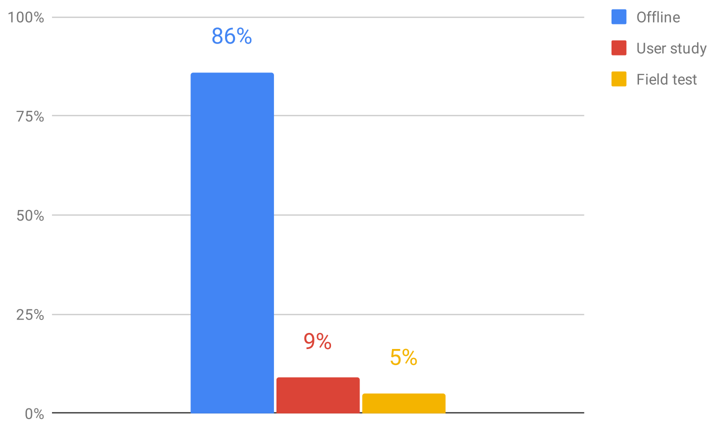

Figure 8: Experiment Types.

target constraint that should be met. And these papers then use computational metrics to assess to what extent an algorithm is able to meet those targets. The target distribution, e.g., of popular and long-tail content, is usually assumed to be given or to be a system parameter. To what extent a certain distribution or metric value would be considered fair by users or other stakeholders in a given domain is usually not discussed. In any practical application, this question is however fundamental, and again the danger exists that research is stuck in an abstraction trap. In a recent work on job recommendations [96], it was for example found that a debiasing algorithm lead to fairer recommendation without a loss in accuracy. A user study then however revealed that participants actually preferred the original system recommendations.

Main Technical Contributions and Algorithmic Approaches. Looking only at the *technical* papers, we identified three main groups of technical contributions: (i) works that report outcomes of data analyses or which compare recommendation outcomes, (ii) works that propose algorithmic approaches to increase the fairness of the recommendations, and (iii) works that propose new metrics or evaluation approaches. Figure 9 shows the distribution of papers according to this categorization.

We observe that most technical papers aim to make the recommendations of a system fairer, e.g., by reducing biases or by aiming to meet a target distribution. Technically, in analogy to context-aware recommender systems [124], this "fairness step" can be done (i) in a pre-processing step, (ii) integrated in the ranking model (modeling approaches), or (iii) in a post-processing step. Figure 10 shows what is common in the current literature. Methods that rely on some form of pre-processing are comparably rare. Typical approaches for modeling approaches include specific fairness-aware loss functions or optimizing methods that consider certain constraints. Post-processing approaches are frequently based on re-ranking.

Overall, the statistics on the one hand point to a possible research gap in terms of works that aim to understanding what leads to unfair recommendations and how severe the problems are for different algorithmic approaches in

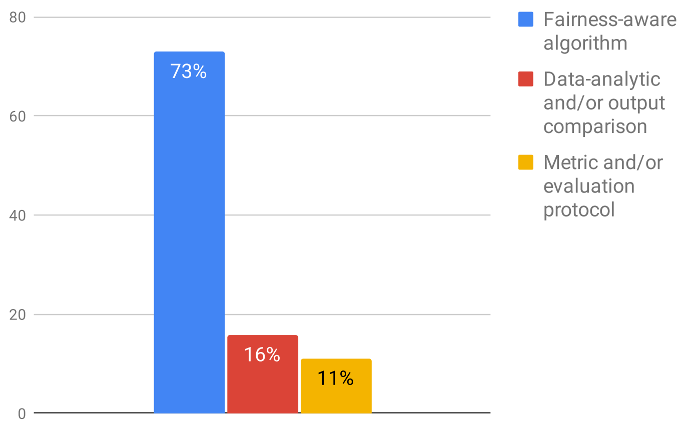

Figure 9: Technical Focus of Papers.

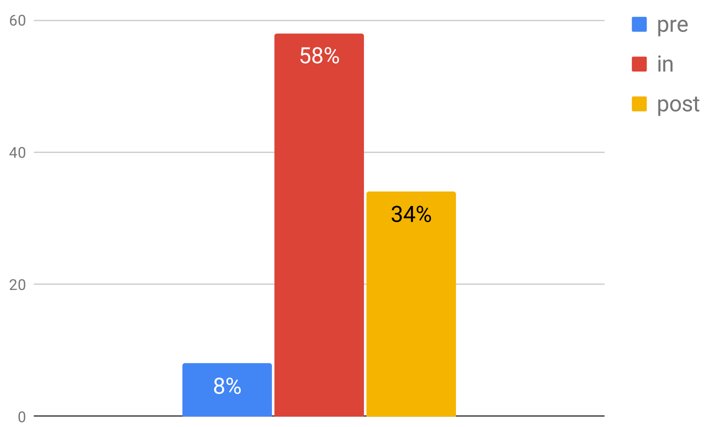

Figure 10: Fairness Step.

particular domains. In the future, it might therefore be important to focus more on analytical research, as advocated also in [101], e.g., to understand the idiosyncrasies of a particular application scenario instead of aiming solely for general-purpose algorithms. On the other hand, the relatively large amount of works that propose new ways of evaluating indicate that the field is not yet mature and has not yet established a standardized research methodology. We discuss evaluation metrics next.

Evaluation Metrics. We found that in offline experiments a wide range of computational metrics is used to assess the fairness of a set of recommendations. The choice of the particular fairness metric mainly depends on the underlying notion of fairness, e.g., if it is about individual or group fairness, or if it is about user or item fairness. In many cases, such fairness metrics are considered in parallel with accuracy metrics based on the assumption that there often is a trade-off between accuracy and fairness.

The following main groups of metrics can be identified.

- *Item exposure based metrics:* These metrics are used when the goal is to establish fairness for items or their providers. The underlying assumption is individual items or groups of items (e.g., long-tail items or musical tracks of female artists) are underrepresented in the recommendations by the system. Typical examples in the studied papers are metrics that determine how many items from the long tail appear in recommendations after making an algorithm fairness-aware [70, 78, 68, 125, 126]. Others consider the coverage of items (e.g., in the form of aggregate diversity) or the Gini index as a metric [63, 79, 78]. In some cases, existing measures, e.g., the average popularity of the recommended items are used [126]; other times specific fairness metrics are proposed. In some cases, these fairness metrics are again based on item popularity, e.g., in [127][5]. Besides item popularity, sometimes protected attributes like gender (of an artist related to the item) are used to determine the different groups for which the exposure should be increased or capped, e.g. [70].
- *Divergence or deviation based metrics:* This class of metrics is used both for user fairness and for item fairness. For user-related fairness, some works consider it fair when a user receives item recommendations which match the characteristics and distributions of items they liked in the past, for example, in terms of the genre distribution in case of movies. Fairness is thus assumed to be achieved when the recommendations are *calibrated* with the past user profile. Correspondingly, measures like the Kullback-Leibler divergence are applied, which are commonly used on the literature on calibration [97, 99, 98]. Recently, bias disparity [43] has become a common alternative for measuring *miscalibration* by looking at how proportions of item categories from input data change in recommendation output [44, 45]. For item-related fairness, similar ideas can be applied and one can compare the distribution of the recommendations with some *target* distribution, which is considered fair. In [35], for example, Generalized Cross Entropy is used to assess the distance of the actual recommendations to the target distribution. Sometimes, the output of a "standard" recommender is taken as target distribution in evaluation of fairness e.g., for a "network-friendly" [129] or a coverage-aware recommender [125]. In some works, also simpler deviation measures are used, e.g., MAD [52, 35, 36] or GAP [76, 130], which both share the characteristics to be point-wise (i.e., non-distributional) and not using a target representation.
- *Recommendation quality measures for groups:* In a number of works on user-related fairness, the goal is to ensure that no group of users is discriminated by receiving recommendations of lower quality than another (privileged or majority) group. To evaluate this aspect, common accuracy measures can be applied and compared across groups, for example, NDCG [117, 35] for user groups, or relevance [68] for item groups.
- *Recommendation quality measures for individuals:* In works that aim at individual fairness—these are mostly works on group recommendation—a common goal often is to make recommendations that are acceptable for all group members, i.e., where the preferences of none of the group members is ignored. As examples of such works, we can name [131, 95, 60, 94, 123]. For instance, in [123] the authors try to ensure that each

[5] Such metrics are thus often highly correlated with other popularity metrics [128].

member likes at least one item in the group list (*proportionality*), or that each user is *envy-free* i.e., there is at least one item for which the user preference is among the highest in the group, so that the user is not envious against other members of the group, who always get a better deal. In [131] fairness is measured by looking at disagreement between group members, over a sequence of recommendations. Another line of research explore individual fairness in a non-group recommendation setting. Such works quantify unfairness on an individual level, by considering recommendation quality per item/user and then aggregating over the entire set of items/users, without dividing them into groups [88, 105].

Table 2 provides a number of examples of metrics used in the surveyed literature. Note that some metrics can be considered as examples for more than one category.

Discussion. A main problem when using computational metrics in offline experiments in general is that it is often unclear to what extent these metrics translate to better systems in practice. In non-fairness research, this typically amounts to the question if higher prediction accuracy on past data will lead to more value for consumers or providers, e.g., in terms of user satisfaction or business-oriented key performance indicators, see [1]. In fairness research, the corresponding questions are if users would actually consider the recommendations fairer or if a fairness-aware algorithm would lead to different behavior of the users. Unfortunately, research that involves humans is very rare. An example for a work that considers the effects of fair rankings can be found in [54], where mixed effects were observed.

Another potential issue of the used metrics is that they may be a strong over-simplification or too strong abstraction of the real problems. Consider the problem of recommending long-tail (less popular) items, which is in the focus of many research works. The metrics we found that measure how many long tail items are recommended usually do not differentiate whether the recommended item is a "good" one or not, by using some form of quality assessment. As mentioned, some items may be unpopular just because of their poor quality. Also, in many of these works it is not clear what a desirable level of exposure of long-tail items would be. This is a problem that is particularly pronounced also for many works that measure fairness through the deviation of the recommendations from some target (desirable) distribution. In technical terms, adjusting the recommendations to be closer to some target distribution can be done with almost trivial and very efficient means like re-ranking. The true and important question however is how we know the target distribution in a given application context.

Generally, we also found a number of works where biased recommendations (e.g., towards popular items) were equated with unfairness. As discussed, this assumption may be too strong. In some of these papers, no deeper discussion is provided why the biases leads to unfairness in a certain application context. The fairness aspect of these papers is therefore often shallow, in parts leading to the impression that fairness was mainly used as a trending label for a work that is mainly about bias mitigation. Similar observations can be made for some papers that argue that calibrating recommendations leads to fairness, which can probably not be safely stated in general.

When considering recommendation quality measures for groups, the assumption is that different groups should have equal recommendation quality (to treat them all alike) or unequal quality (e.g., because some group has paid for

Table 2: Overview of evaluation metrics and its corresponding fairness notion that is aiming to capture.

Category of Metrics	Examples
Item exposure based metrics	Popularity Bias [77, 36] Popularity Count [77] Exposure variance [105] Exposure/visibility gain [57] Exposure/relevance ratio [105] Weighted proportional fairness [73] Fraction of satisfied producers, inequality in exposures distribution [104] Popularity rate (and long-tail rate) [78] Gini index of: recommendation frequency [63]; Popularity scores [79] Ranking-based statistical parity, ranking-based equal-opportunity [127]
Divergence or deviation based metrics	KL divergence of exposure, normalized discounted KL divergence [72] Bias disparity [43, 44, 45] Generalized Cross Entropy [35] Max individual deviation, total variation distance, and KL-divergence (between distributions of item interactions) [129] Mean Absolute Difference (MAD) [52, 35, 36] Group Average Popularity (GAP) [76, 130] User/item deviation cost [125] Disparity of exposure [68]
Recommendation quality measures for groups	NDCG (user groups) [117, 59] F1 score [59] F-statistic [41] Relevance (item groups) [68] MAD-ranking [35]
Recommendation quality measures for individuals	• **Group recommendation setting:** Mean, Minimum and Min-Max ratio of NDCG and Average Relevance; Pearson correlation, MAE [95] Average Reciprocal Hit Rank [60] Zero-recall; Mean, Minimum and Min-Max ratio of: Recall, Discounted First Hit, and NDCG [94] Mean and Min-Max difference of user satisfaction [131] • **Classical recommendation setting:** User NDCG variance [105] User MSE variance [85] Item Statistical Parity and Item Equal Opportunity [88] Mean pairwise disparity of item exposure and relevance [132] Mean Discounted Gain (MDG) for worst items [82]

a better service). As argued above, in most applications of recommenders the recommendations will be better in terms of accuracy measures for active users than for less active users. Some papers in this survey consider this unfair, but

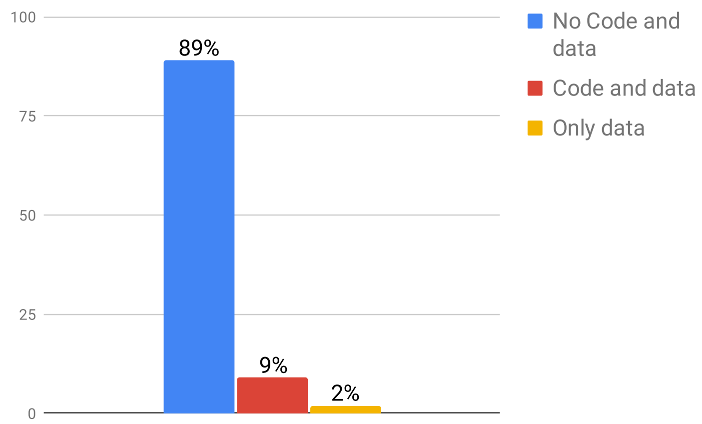

Figure 11: Level of Reproducibility (Shared Artifacts).

this line of argumentation is not easy to follow. Certainly, there may be scenarios where there are particular protected attributes for which it may be desirable to not have largely varying accuracy levels across the groups. In many of the surveyed papers no realistic use cases are however given.

Finally, looking at individual fairness in group recommendation scenarios, a multitude of aggregation strategies were proposed over the years such as Least Misery or Borda Count [92]. The literature on group recommender systems—which is now revived under the term fairness—however does not provide a clear conclusion regarding which aggregation metric should be used in a given application. Also in this area researchers have be stuck in an abstraction trap and more (multi-disciplinary) research seems required to understand group recommendation processes, see [133] for an observational study in the tourism domain.

Reproducibility. The lack of reproducibility can be a major barrier to achieve progress in AI [134], and recent studies indicate that limited reproducibility is a substantial issue also in recommender systems research [135, 136]. Figure 11 shows for how many of the studied *technical* papers, artifacts were shared to ensure reproducibility of the reported experiments. While the level of reproducibility seems to be higher than in general AI [134], still for the large majority of the considered works authors did not share any code or data.

4.6. Landscape Overview

Fairness is a multi-faceted subject. In order to provide an encompassing understanding of different fairness dimensions, we have developed an simplified taxonomy and landscape of fairness research in recommender systems, as shown in Figure 12. The landscape's main aspects can be summarized based on the following questions.

- **How is fairness implemented?** Depending on which step of the recommendation pipeline we change, fairness-enhancing systems can be divided into are pre-, in- and post-processing techniques. Here we also note that the main patterns are in- and post-processing (typically re-ranking), probably due to the advantage of an easier applicability to existing systems.

- **What is the target representation?** The *target representation* is defined as the ideal representation (i.e., proportion or distribution of exposure) [69]. In other works, this is also referred to as *target distribution* (of benefits such as exposure or relevance). Even though this aspect has not been specifically analyzed in the previously presented figures, we have identified three main target representations against which most fairness metrics compare: catalog size, relevance, and parity. These representations match those introduced in [69], where authors state that the choice of the representation target depends on the application domain. Among these, the most common interpretation is that items should be recommended equally for each group, hence, using a parity-based representation target. However, there are also other aspects and fairness notions that do not use this assumption, as discussed in Section 2.3.
- **What is the benefit of fairness?** As in the previous case, for the sake of conciseness, we have not considered this dimension in this detailed analysis, but it is worth mentioning that fairness definitions can be categorized depending on whether its main benefit is based on exposure (by assessing if items are exposed in a uniform or fair way) or relevance (with the additional constraint on the exposure that it must be effective, that is, it should match the user preferences). In principle, any information seeking system (such as search engines or recommender systems) should aim for relevance-based benefits. However, considering the difficulty of these tasks, by measuring and achieving a situation with fair exposure, the subsequent measurements on the system would already be impacted and improved, from a fairness perspective and, hence, it is a reasonable goal to obtain.
- **How is fairness measured?** Fairness evaluation, as any other experimental research, can be performed through qualitative or quantitative methods. As discussed in Section 4.5, qualitative approaches are currently almost never taken, and most of the analyses are done by quantitative approaches such as offline experiments or A/B tests.
- **On which level is fairness considered?** Fairness can be defined on a group level or individual level, as discussed above. Today, group-level fairness is the prevalent option, most likely because measuring (operationalizing) group fairness is easier than individual fairness. In other words, what it means for two individuals to be similar is task-sensitive and more difficult than segmenting users/items into groups based on a sensitive feature, as is often done in the examined literature of group fairness. This might also have social implications, as many major considerations of fairness in the literature, including gender equality, demographic equality, and others, are predicated on the concept of group fairness. It is important to note that the primary limitation of group fairness is the decreasing reliability of sensitive attributes in recent years due to privacy concerns and firms' reluctance to share such information.
- **Fairness for whom?** In many cases, the circumstance for making a recommendation is intrinsically multi-sided. As a result, any of the *stakeholders* engaged, as well as the platform itself, may be affected by (un)fairness.

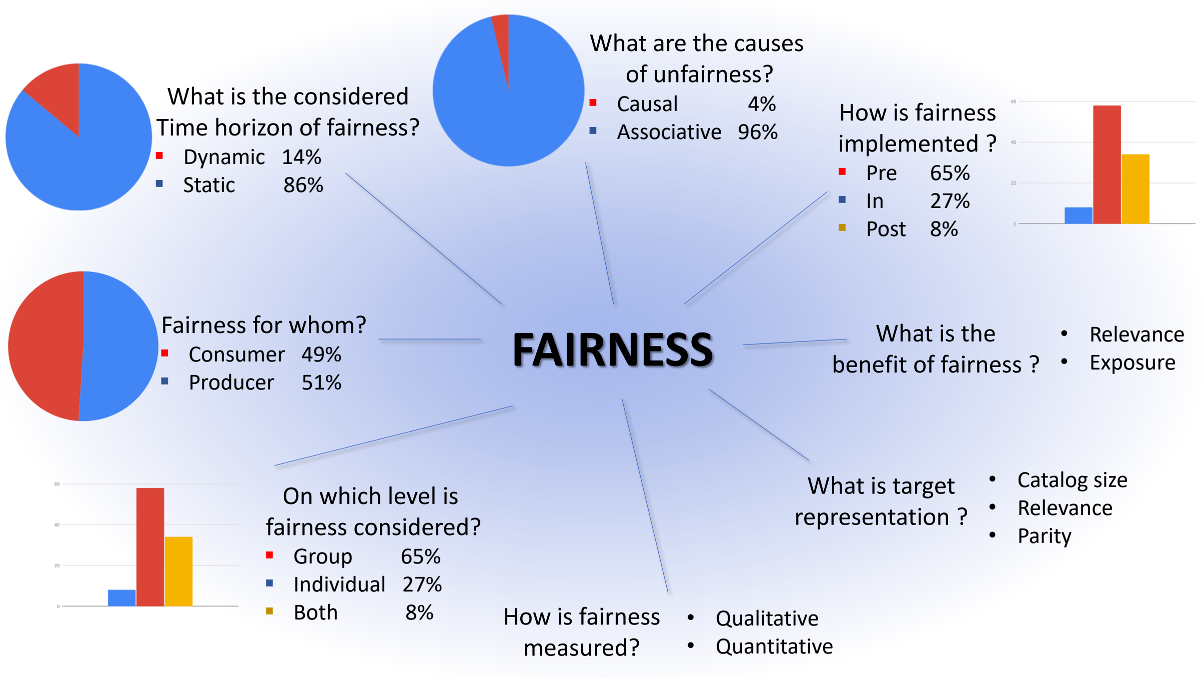

Figure 12: Taxonomy and Landscape.

Through our survey, we found that there is a balance in the literature between consumer and provider viewpoints.

- **What is the considered time horizon of fairness?** Fairness can be pursued in a static way (or: *one-shot*), or dynamically over time, taking into account shifts in the item catalog, user tastes, etc. However, practically we observe a prevalence of the former, with the latter including new trends like reinforcement learning-based approaches.
- **What are the causes of unfairness?** The dominant pattern of fairness-enhancing approaches seems to pursue a static, associative, group-level notion of fairness, inheriting from fair ML traditional research. Hence, papers considering relatively new approaches such as causal inference and long-term fairness are more rare. We can describe this as a research gap, i.e., there should be more research into the reasons of unfairness through the lens of causality and counterfactuals.

5. Discussion

Summary of Main Observations. Due to today's broad and increasing use of AI in practical applications, questions relating to potential harms of AI-powered systems have received more and more attention in recent years, both in academic research, tech industry, and within political organizations. Fairness is often considered a central component of what is sometimes called *responsible* AI. These developments can also be seen in the area of recommender systems, where we observed a strong increase in terms of publications on fairness since the mid-2010s, cf. Figure 1.

Looking closer at the research contributions from the field of computer science, we observe that the large majority of works aim to provide technical solutions, and that the technical contributions are predominantly fairness-aware algorithms (cf. Figure 2 and Figure 9). In contrast, only comparably limited research activity seems to take place on topics that go beyond the computer science perspective. While algorithmic research is certainly important, focusing almost exclusively on improving algorithms in terms of optimizing an abstract computational fairness metric may be too limited. Ultimately, however, our goal should rather be to design "*algorithmic systems that support human values*" [137] and avoid potential abstraction traps, similar as in the general area of fair ML.

On the positive side, we find that researchers in fair RS are addressing various notions of fairness (cf. Figures 3 to 6), e.g., they deal with questions both of individual fairness and of group fairness. In addition, the community has expanded the scope of fairness considerations beyond some affected humans and has developed various approaches to deal with fairness towards items and providers. This is different from many other traditional application areas of fair ML, e.g., credit default prediction, where the affected humans are usually the main focus of research.

Looking at the considered application domains and datasets, we observe that various domains are addressed. However, the large majority of technical papers report experiments with datasets from the media domain (videos and music), cf. Figure 7. Specifically, some of the MovieLens datasets are frequently used either as a concrete use case or as a way to at least provide reproducible results, given that the set of fairness aspects that can be reasonably studied with such datasets seems limited. All in all, there seems to be a certain lack of real-world datasets for real-world fairness problems, which is why researchers frequently also rely on synthetic data or on protected groups that are artificially introduced into a given recommendation dataset.

In terms of the research methodology, offline experiments using the described datasets are the method of choice for most researchers, cf. Figure 8. Only very few works rely on studies that have the human in the loop, *which points to a major research gap in fair recommender systems*. In the context of these offline evaluations, a rich variety of evaluation approaches and computational metrics are used. The way the research problems are operationalized however often appears to be an oversimplification of the underlying problem. In many research works, for example, (popularity) biases are equated with unfairness, which we believe is not necessarily the case in general. Some of the surveyed works also seem to "re-brand" existing research on beyond-accuracy quality aspects of recommendations—e.g., on diversity or calibration—as fairness research, sometimes without providing a plausible and realistic use case. Finally, in almost all works some "gold standard" for fair recommendations is assumed to be given, e.g., in the form of a target distribution regarding item exposures. With the goal of providing generic algorithmic solutions, little or no guidance is however usually provided on how to decide or determine this gold standard for a given use case. While general-purpose solutions are certainly desirable, the danger of being stuck in abstraction trap with limited practical impact increases.

Future Directions. Our analysis of the current research landscape points to a number of further research gaps. Considering the type of contributions and the different notions of fairness, we find that today's research efforts are not balanced. Most published works are algorithmic contributions and use offline evaluations with a variety of proxy metrics to assess fairness. Moreover, these offline evaluations are based on one particular point in time. As such, these evaluations do not consider longitudinal dynamics that may emerge (a) when the fairness goals change over time or (b) when an algorithm's output changes over time, e.g., when a fairness intervention gradually improves the recommendations. This limitation of static offline evaluations also becomes more acknowledged in the general recommender systems literature. Simulation approaches are recently often considered as one promising approach to model such longitudinal dynamics [11, 106, 107, 108]. Causal models, in contrast to associative ones, also received very limited research attention so far.

Through our survey, we furthermore identified a number of promising research problems for which only few works exist so far:

- **Challenge 1:** *Achieving a realistic and useful definition for fairness.* As discussed before, there are several definitions for fairness, not only in the RS literature but in ML and AI in general. This provokes incompatibility between some of these definitions and potential disagreement, where one metric may conclude that a recommender system is fair and another the opposite. As a consequence, it is not easy to find a proper balance between different notions of fairness and the performance of the recommendation models. An example of a respective proposal can be found in [73],where the authors employ metrics that capture the cumulative reward in a way that combines accuracy and fairness while aiming to improve both.

  However, this is not the only problem we have identified in our literature review. As stated in Section 4.5, the seldom use of user studies and field tests make it impossible to incorporate the user perception into our understanding of what should be defined as a fair recommendation. In fact, some works propose to move from notions of equality to those of equity and independence [35], but even these general definitions that may work at a societal level, may not necessarily make sense depending on the domain or the user needs.

  One interesting approach to address these issues is through root cause analysis via causal modeling. This technique is sometimes used to detect anomalies so that not only the outlier data is captured but the reason for such behavior is identified [138]. By translating this concept into recommendation, the underlying cause of the unfairness may be explained, and the user (or the system developer) may provide feedback and understand if those reasons are appropriate and if they correspond with a situation that is actually unfair.

- **Challenge 2:** *Building on appropriate data to assess fairness.* As discussed in Section 4.4, some datasets used in the literature do not contain sensitive attributes at all. This problem has been addressed in different ways, none of them perfect but fruitful towards the goal of mimicking the evaluation of recommender systems in realistic scenarios. A first possibility is to perform data augmentation, where the main idea is, without changing the underlying data and algorithm, to be able to remove biases from the data to provide higher-quality information to the algorithms [85]. Another, more popular, possibility is to use of simulation instead of real-world datasets. Various recent papers use simulation and synthetic data to evaluate fairness in search scenarios, see [46]. This may require more advanced techniques in the evaluation step, such as counterfactual evaluation, in order to properly interpret the data coming from A/B logged interactions once interventions have been performed through a recommendation algorithm, for example, by focusing on improving item exposure [139].

- **Challenge 3:** *Understanding fairness in reciprocal settings.* Maintaining the utility of stakeholders in reciprocal settings is a new notion of fairness [71], even though reciprocal recommender systems have been studied (although not as frequently as other systems) in the past and remain at the core of social network and matching platforms [140]. In that latter work, the authors define fairness as an equilibrium between parties where there are 'buyers' and 'sellers' and each seller has the same value or 'price'; hence, in their notion of "Walrasian Equilibrium" they are treated fairly by considering at the same time (a) the disparity of service, (b) the similarity of mutual preference, and (c) the equilibrium of demand and supply.
  By considering the importance of this type of systems, being able to operationalize a reasonable definition for this context is foreseen as a major challenge to tackle in the future.
- **Challenge 4:** *Fairness auditing.* As stated in [141], algorithm auditing is the research and practice of assessing, mitigating, and assuring an algorithm's legality, ethics, and safety. In that work, the authors consider bias and discrimination as one of the main verticals of algorithm auditing. Hence, auditing recommender systems should become a priority in the near future, and the fairness dimension is, by definition, one of the most important aspects to be consider in that process. As an example, we want to hightlight that the authors from [31] aimed at auditing decision making systems, but faced important issues since their agents were banned from the platform that was meant to be analyzed (Facebook NewsFeed). Hence, there are technical difficulties that may make this challenge even harder to achieve, despite its importance in legal and ethical dimensions.

Finally, one main fundamental problem of current research on fair recommender systems is that it is not entirely clear yet how impactful it is in practice. Algorithmic research is too often based on a very abstract and probably overly simplistic operationalization of the research problem, using computational metrics for which it is not clear if they are good proxies for fairness in a particular problem setting. In such a research approach, fundamental questions of what is a fair recommendation in a given situation are not discussed. Correspondingly, the choice of application domains sometimes seems arbitrary (based on dataset availability), and the fairness challenges often appear almost artificial. Moreover, connections to existing works and theories developed in the social sciences are rarely established in the published literature, and fairness is often simply treated as an algorithmic problem, e.g., to make recommendations that match a pre-defined target distribution. In some ways, current research shares challenges with many works in the area of Explainable AI, where many insights from social sciences exist, and where it is often neglected that explainable AI, like recommendation, to a large extent is a problem of human-computer interaction [142]. As a consequence, much more fundamental research on fairness, its definition in a given problem setting, and its perception by the involved stakeholders is needed. This, in turn, requires a multidisciplinary approach, involving not only researchers from different areas of computer sciences, but also including subject-matter experts from real-world problem settings and maybe scholars from fields outside computer science.

References

[1] D. Jannach, M. Jugovac, Measuring the business value of recommender systems, ACM TMIS 10 (4) (2019) 1–23.

[2] C. Trattner, D. Jannach, E. Motta, I. C. Meijer, N. Diakopoulos, M. Elahi, A. L. Opdahl, B. Tessem, N. Borch, M. Fjeld, L. Øvrelid, K. D. Smedt, H. Moe, Responsible Media Technology and AI: Challenges and Research Directions, AI and Ethics 2022 (2022).

[3] R. Burke, Multisided fairness for recommendation, in: 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017), 2017.

[4] M. D. Ekstrand, A. Das, R. Burke, F. Diaz, Fairness and discrimination in information access systems, CoRR abs/2105.05779 (2021).

[5] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2021).

[6] S. Barocas, M. Hardt, A. Narayanan, Fairness and Machine Learning, fairmlbook.org, 2019, http://www.fairmlbook.org.

[7] H. Abdollahpouri, G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, L. Pizzato, Multistakeholder recommendation: Survey and research directions, User Modeling and User-Adapted Interaction 30 (2020) 127–158.

[8] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and abstraction in sociotechnical systems, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, 2019, p. 59–68.

[9] A. Ashokan, C. Haas, Fairness metrics and bias mitigation strategies for rating predictions, Inf. Process. Manag. 58 (5) (2021) 102646.

[10] D. Jannach, L. Lerche, I. Kamehkhosh, M. Jugovac, What recommenders recommend: an analysis of recommendation biases and possible countermeasures, User Modeling and User-Adapted Interaction 25 (5) (2015) 427–491.

[11] N. Ghanem, S. Leitner, D. Jannach, Balancing consumer and business value of recommender systems: A simulation-based analysis, arXiv preprint arXiv:2203.05952 (2022).

[12] D. Jannach, G. Adomavicius, Price and profit awareness in recommender systems, in: Proceedings of the ACM RecSys 2017 Workshop on Value-Aware and Multi-Stakeholder Recommendation, 2017.

[13] Y. R. Shrestha, Y. Yang, Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems, Algorithms 12 (9) (2019) 199.

[14] E. Pitoura, K. Stefanidis, G. Koutrika, Fairness in rankings and recommendations: An overview, CoRR abs/2104.05994 (2021).

[15] M. Zehlike, K. Yang, J. Stoyanovich, Fairness in ranking: A survey, arXiv preprint arXiv:2103.14000 (2021).

[16] S. Verma, J. Rubin, Fairness definitions explained, in: Y. Brun, B. Johnson, A. Meliou (Eds.), Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018, ACM, 2018, pp. 1–7.

[17] S. Verma, R. Gao, C. Shah, Facets of fairness in search and recommendation, in: L. Boratto, S. Faralli, M. Marras, G. Stilo (Eds.), Bias and Social Aspects in Search and Recommendation - First International Workshop, BIAS 2020, Lisbon, Portugal, April 14, 2020, Proceedings, Vol. 1245 of Communications in Computer and Information Science, Springer, 2020, pp. 1–11.

[18] D. K. Mulligan, J. A. Kroll, N. Kohli, R. Y. Wong, This thing called fairness: Disciplinary confusion realizing a value in technology, Proc. ACM Hum. Comput. Interact. 3 (CSCW) (2019) 119:1–119:36.

[19] M. Elahi, D. Jannach, L. Skjærven, E. Knudsen, H. Sjøvaag, K. Tolonen, Øyvind Holmstad, I. Pipkin, E. Throndsen, A. Stenbom, E. Fiskerud, A. Oesch, L. Vredenberg, C. Trattner, Towards responsible media recommendation, AI and Ethics forthcoming (2021).

[20] M. D. Ekstrand, A. Das, R. Burke, F. Diaz, Fairness and discrimination in information access systems, arXiv preprint arXiv:2105.05779 (2021).

[21] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, X. He, Bias and debias in recommender system: A survey and future directions, CoRR abs/2010.03240 (2020).

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

[23] B. Friedman, H. Nissenbaum, Bias in computer systems, ACM Trans. Inf. Syst. 14 (3) (1996) 330–347.

[24] B. Kitchenham, Procedures for performing systematic reviews, Keele University. Technical Report TR/SE-0401, Department of Computer Science, Keele University, UK (2004).

[25] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, The connection between popularity bias, calibration, and fairness in recommendation, in: Fourteenth ACM Conference on Recommender Systems, 2020, p. 726–731.

[26] D. Pedreshi, S. Ruggieri, F. Turini, Discrimination-aware data mining, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’08, 2008, p. 560–568.

[27] D. Jannach, P. Resnick, A. Tuzhilin, M. Zanker, Recommender systems - beyond matrix completion, Communications of the ACM 59 (11) (2016) 94–102.

[28] D. Jannach, P. Pu, F. Ricci, M. Zanker, Recommender systems: Past, present, future, AI Magazine 42 (3) (2021) 3–6.

[29] L. Schelenz, Diversity-aware recommendations for social justice? exploring user diversity and fairness in recommender systems, in: Adjunct Publication of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, 2021, pp. 404–410.

[30] N. Sonboli, J. J. Smith, F. Cabral Berenfus, R. Burke, C. Fiesler, Fairness and Transparency in Recommendation: The Users’ Perspective, 2021, p. 274–279.

[31] T. D. Krafft, M. P. Hauer, K. A. Zweig, Why do we need to be bots? what prevents society from detecting biases in recommendation systems, in: Bias and Social Aspects in Search and Recommendation - First International Workshop, BIAS 2020, Vol. 1245, 2020, pp. 27–34.

[32] Y. Li, Y. Ge, Y. Zhang, Tutorial on Fairness of Machine Learning in Recommender Systems, 2021, p. 2654–2657.

[33] R. Binns, On the apparent conflict between individual and group fairness, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* ’20, 2020, p. 514–524.

[34] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R. S. Zemel, Fairness through awareness, in: S. Goldwasser (Ed.), Innovations in Theoretical Computer Science 2012, 2012, pp. 214–226.

[35] Y. Deldjoo, V. W. Anelli, H. Zamani, A. Bellogin, T. Di Noia, A flexible framework for evaluating user and item fairness in recommender systems, User Modeling and User-Adapted Interaction (2021) 1–47.

[36] Y. Deldjoo, A. Bellogin, T. Di Noia, Explaining recommender systems fairness and accuracy through the lens of data characteristics, Information Processing & Management 58 (5) (2021) 102662.

[37] C. Wu, F. Wu, X. Wang, Y. Huang, X. Xie, Fairness-aware news recommendation with decomposed adversarial learning, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, 2021, pp. 4462–4469.

[38] S. Gorantla, A. Deshpande, A. Louis, On the problem of underranking in group-fair ranking, in: International Conference on Machine Learning, PMLR, 2021, pp. 3777–3787.

[39] X. Wang, N. Thain, A. Sinha, F. Prost, E. H. Chi, J. Chen, A. Beutel, Practical compositional fairness: Understanding fairness in multi-component recommender systems, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 2021, pp. 436–444.

[40] A. Ghosh, R. Dutt, C. Wilson, When fair ranking meets uncertain inference, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1033–1043.

[41] M. Wan, J. Ni, R. Misra, J. McAuley, Addressing marketing bias in product recommendations, in: Proceedings of the 13th International Conference on Web Search and Data Mining, 2020, pp. 618–626.

[42] B. Edizel, F. Bonchi, S. Hajian, A. Panisson, T. Tassa, Fairecsys: mitigating algorithmic bias in recommender systems, International Journal of Data Science and Analytics 9 (2) (2020) 197–213.

[43] V. Tsintzou, E. Pitoura, P. Tsaparas, Bias disparity in recommendation systems, in: Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Vol. 2440, 2019.

[44] M. Mansoury, B. Mobasher, R. Burke, M. Pechenizkiy, Bias disparity in collaborative recommendation: Algorithmic evaluation and comparison, in: Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), 2019.

[45] K. Lin, N. Sonboli, B. Mobasher, R. Burke, Crank up the volume: Preference bias amplification in collaborative recommendation, in: Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Rec-

ommender Systems (RecSys 2019), Vol. 2440 of CEUR Workshop Proceedings, 2019.

[46] S. C. Geyik, S. Ambler, K. Kenthapadi, G. Karypis, 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, KDD 2019, 2019, pp. 2221–2231.

[47] B. Xia, J. Yin, J. Xu, Y. Li, We-rec: A fairness-aware reciprocal recommendation based on walrasian equilibrium, Knowl. Based Syst. 182 (2019).

[48] R. Burke, N. Sonboli, A. Ordonez-Gauger, Balanced neighborhoods for multi-sided fairness in recommendation, in: Conference on Fairness, Accountability and Transparency, FAT 2018, Vol. 81 of Proceedings of Machine Learning Research, 2018, pp. 202–214.

[49] A. Chakraborty, J. Messias, F. Benevenuto, S. Ghosh, N. Ganguly, K. P. Gummadi, Who Makes Trends? Understanding Demographic Biases in Crowdsourced Recommendations, in: Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, 2017, pp. 22–31.

[50] S. Gorantla, A. Deshpande, A. Louis, On the problem of underranking in group-fair ranking, in: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Vol. 139 of Proceedings of Machine Learning Research, 2021, pp. 3777–3787.

[51] Y. Zheng, T. Dave, N. Mishra, H. Kumar, Fairness in reciprocal recommendations: A speed-dating study, in: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP 2018, 2018, pp. 29–34.

[52] Z. Zhu, X. Hu, J. Caverlee, Fairness-aware tensor-based recommendation, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, 2018, pp. 1153–1162.

[53] J. Bobadilla, R. Lara-Cabrera, Á. González-Prieto, F. Ortega, Deepfair: Deep learning for improving fairness in recommender systems, Int. J. Interact. Multim. Artif. Intell. 6 (6) (2021) 86–94.

[54] T. Sühr, S. Hilgard, H. Lakkaraju, Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring, 2021, p. 989–999.

[55] A. B. Melchiorre, N. Rekabsaz, E. Parada-Cabaleiro, S. Brandl, O. Lesota, M. Schedl, Investigating gender fairness of recommendation algorithms in the music domain, Inf. Process. Manag. 58 (5) (2021) 102666.

[56] L. Weydemann, D. Sacharidis, H. Werthner, Defining and measuring fairness in location recommendations, in: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-based Recommendations, Geosocial Networks and Geoadvertising, LocalRec@SIGSPATIAL 2019, 2019, pp. 6:1–6:8.

[57] E. Gómez, C. Shui Zhang, L. Boratto, M. Salamó, M. Marras, The Winner Takes It All: Geographic Imbalance and Provider (Un)Fairness in Educational Recommender Systems, 2021, p. 1808–1812.

[58] Q. Hao, Q. Xu, Z. Yang, Q. Huang, Pareto optimality for fairness-constrained collaborative filtering, in: H. T. Shen, Y. Zhuang, J. R. Smith, Y. Yang, P. Cesar, F. Metze, B. Prabhakaran (Eds.), MM '21: ACM Multimedia Conference, ACM, 2021, pp. 5619–5627.

[59] Y. Li, H. Chen, Z. Fu, Y. Ge, Y. Zhang, User-oriented fairness in recommendation, in: Proceedings of The Web Conference 2021, WWW '21, 2021, p. 624–632.

[60] Y. Xiao, Q. Pei, L. Yao, S. Yu, L. Bai, X. Wang, An enhanced probabilistic fairness-aware group recommendation by incorporating social activeness, J. Netw. Comput. Appl. 156 (2020) 102579.

[61] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, S. Xu, S. Geng, C. Shah, Y. Zhang, G. de Melo, Fairness-aware explainable recommendation over knowledge graphs, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, 2020, pp. 69–78.

[62] A. Chakraborty, G. K. Patro, N. Ganguly, K. P. Gummadi, P. Loiseau, Equality of voice: Towards fair representation in crowdsourced top-k recommendations, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, 2019, pp. 129–138.

[63] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, E. C. Malthouse, User-centered evaluation of popularity bias in recommender systems, in: Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, UMAP 2021, ACM, 2021, pp. 119–129.

[64] N. N. Htun, E. Lecluse, K. Verbert, Perception of fairness in group music recommender systems, in: 26th International Conference on

Intelligent User Interfaces, 2021, p. 302–306.

[65] C. Lin, X. Liu, G. Xv, H. Li, Mitigating sentiment bias for recommender systems, in: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 31–40.

[66] A. Gharahighehi, C. Vens, K. Pliakos, Fair multi-stakeholder news recommender system with hypergraph ranking, Inf. Process. Manag. 58 (5) (2021) 102663.

[67] A. Ferraro, Music cold-start and long-tail recommendation: bias in deep representations, in: T. Bogers, A. Said, P. Brusilovsky, D. Tikk (Eds.), Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, 2019, pp. 586–590.

[68] L. Boratto, G. Fenu, M. Marras, Interplay between upsampling and regularization for provider fairness in recommender systems, User Model. User Adapt. Interact. 31 (3) (2021) 421–455.

[69] Ö. Kirnap, F. Diaz, A. Biega, M. D. Ekstrand, B. Carterette, E. Yilmaz, Estimation of fair ranking metrics with incomplete judgments, in: WWW '21: The Web Conference 2021, 2021, pp. 1065–1075.

[70] D. Shakespeare, L. Porcaro, E. Gómez, C. Castillo, Exploring artist gender bias in music recommendation, in: Proceedings of the Workshops on Recommendation in Complex Scenarios and the Impact of Recommender Systems co-located with 14th ACM Conference on Recommender Systems (RecSys 2020), Vol. 2697 of CEUR Workshop Proceedings, 2020.

[71] B. Xia, J. Yin, J. Xu, Y. Li, We-rec: A fairness-aware reciprocal recommendation based on walrasian equilibrium, Knowledge-Based Systems 182 (2019) 104857.

[72] A. Dash, A. Chakraborty, S. Ghosh, A. Mukherjee, K. P. Gummadi, When the umpire is also a player: Bias in private label product recommendations on e-commerce marketplaces, in: FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 873–884.

[73] W. Liu, F. Liu, R. Tang, B. Liao, G. Chen, P. Heng, Balancing between accuracy and fairness for interactive recommendation with reinforcement learning, in: Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Vol. 12084, 2020, pp. 155–167.

[74] Q. Dong, S. Xie, W. Li, User-item matching for recommendation fairness, IEEE Access 9 (2021) 130389–130398.

[75] D. C. da Silva, M. G. Manzato, F. A. Durão, Exploiting personalized calibration and metrics for fairness recommendation, Expert Systems with Applications 181 (2021) 115112.

[76] B. D. Wundervald, Cluster-based quotas for fairness improvements in music recommendation systems, Int. J. Multim. Inf. Retr. 10 (1) (2021) 25–32.

[77] R. Borges, K. Stefanidis, On mitigating popularity bias in recommendations via variational autoencoders, in: SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, 2021, pp. 1383–1389.

[78] Y. Ge, S. Liu, R. Gao, Y. Xian, Y. Li, X. Zhao, C. Pei, F. Sun, J. Ge, W. Ou, Y. Zhang, Towards long-term fairness in recommendation, in: WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, 2021, pp. 445–453.

[79] W. Sun, S. Khenissi, O. Nasraoui, P. Shafto, Debiasing the human-recommender system feedback loop in collaborative filtering, in: Companion of The 2019 World Wide Web Conference, WWW 2019, ACM, 2019, pp. 645–651.

[80] H. Abdollahpouri, M. Mansoury, R. Burke, B. Mobasher, The unfairness of popularity bias in recommendation, in: Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Vol. 2440, 2019.

[81] Q. Zhu, A. Zhou, Q. Sun, S. Wang, F. Yang, FMSR: A fairness-aware mobile service recommendation method, in: 2018 IEEE International Conference on Web Services, ICWS 2018, San Francisco, CA, USA, July 2-7, 2018, IEEE, 2018, pp. 171–178.

[82] Z. Zhu, J. Kim, T. Nguyen, A. Fenton, J. Caverlee, Fairness among new items in cold start recommender systems, in: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, 2021, pp. 767–776.

[83] Y. Deldjoo, V. W. Anelli, H. Zamani, A. B. Kouki, T. D. Noia, Recommender systems fairness evaluation via generalized cross entropy, in: R. Burke, H. Abdollahpouri, E. C. Malthouse, K. P. Thai, Y. Zhang (Eds.), Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Copenhagen, Denmark,

September 20, 2019, Vol. 2440 of CEUR Workshop Proceedings, CEUR-WS.org, 2019.

[84] Q. Zhu, Q. Sun, Z. Li, S. Wang, FARM: A fairness-aware recommendation method for high visibility and low visibility mobile apps, IEEE Access 8 (2020) 122747–122756.

[85] B. Rastegarpanah, K. P. Gummadi, M. Crovella, 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, 2019, pp. 231–239.

[86] M. Naghiaei, H. A. Rahmani, Y. Deldjoo, Cpfair: Personalized consumer and producer fairness re-ranking for recommender systems, in: SIGIR'22, 2022.

[87] H. A. Rahmani, Y. Deldjoo, A. Tourani, M. Naghiaei, The unfairness of active users and popularity bias in point-of-interest recommendation, arXiv preprint arXiv:2202.13307 (2022).

[88] L. Boratto, G. Fenu, M. Marras, Connecting user and item perspectives in popularity debiasing for collaborative recommendation, Information Processing & Management 58 (1) (2021) 102387.

[89] A. Gupta, E. Johnson, J. Payan, A. K. Roy, A. Kobren, S. Panda, J.-B. Tristan, M. Wick, Online post-processing in rankings for fair utility maximization, in: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21, 2021, p. 454–462.

[90] Z. Zhao, J. Chen, S. Zhou, X. He, X. Cao, F. Zhang, W. Wu, Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation, CoRR abs/2109.07946 (2021).

[91] B. Edizel, F. Bonchi, S. Hajian, A. Panisson, T. Tassa, FaiRecSys: mitigating algorithmic bias in recommender systems, International Journal of Data Science and Analytics (2019) 1–17.

[92] J. Masthoff, Group recommender systems: Combining individual models, in: F. Ricci, L. Rokach, B. Shapira, P. Kantor (Eds.), Recommender Systems Handbook, Springer, 2011.

[93] A. Felfernig, L. Boratto, M. Stettinger, M. Tkali, Group Recommender Systems: An Introduction, Springer, 2018.

[94] M. Kaya, D. Bridge, N. Tintarev, Ensuring Fairness in Group Recommendations by Rank-Sensitive Balancing of Relevance, 2020, p. 101–110.

[95] L. Malecek, L. Peska, Fairness-Preserving Group Recommendations With User Weighting, 2021, p. 4–9.

[96] C. Wang, K. Wang, A. Bian, R. Islam, K. N. Keya, J. R. Foulds, S. Pan, Bias: Friend or foe? user acceptance of gender stereotypes in automated career recommendations, CoRR abs/2106.07112 (2021).

[97] J. Oh, S. Park, H. Yu, M. Song, S.-T. Park, Novel recommendation based on personal popularity tendency, in: ICDM '11, 2011, pp. 507–516.

[98] H. Steck, Calibrated recommendations, in: Proceedings of the 12th ACM Conference on Recommender Systems, 2018, pp. 154–162.

[99] M. Jugovac, D. Jannach, L. Lerche, Efficient optimization of multiple recommendation quality factors according to individual user tendencies, Expert Systems With Applications 81 (2017) 321–331.

[100] B. Xiao, I. Benbasat, E-commerce product recommendation agents: Use, characteristics, and impact, MIS Quarterly 31 (1) (2007) 137–209.

[101] D. Jannach, C. Bauer, Escaping the McNamara Fallacy: Towards more Impactful Recommender Systems Research, AI Magazine 41 (4) (2020) 79–95.

[102] R. Burke, Multisided fairness for recommendation, arXiv preprint arXiv:1707.00093 (2017).

[103] H. Abdollahpouri, R. Burke, Multi-stakeholder recommendation and its connection to multi-sided fairness, in: Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments co-located with the 13th ACM Conference on Recommender Systems (RecSys 2019), Vol. 2440 of CEUR Workshop Proceedings, 2019.

[104] G. K. Patro, A. Biswas, N. Ganguly, K. P. Gummadi, A. Chakraborty, Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms, in: WWW '20: The Web Conference 2020, 2020, pp. 1194–1204.

[105] Y. Wu, J. Cao, G. Xu, Y. Tan, TFROM: A two-sided fairness-aware recommendation model for both customers and providers, in: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2021, pp. 1013–1022.

[106] D. Rohde, S. Bonner, T. Dunlop, F. Vasile, A. Karatzoglou, Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising, arXiv preprint arXiv:1808.00720 (2018).

[107] M. Mladenov, C. Hsu, V. Jain, E. Ie, C. Colby, N. Mayoraz, H. Pham, D. Tran, I. Vendrov, C. Boutilier, RecSim NG: toward principled

uncertainty modeling for recommender ecosystems, CoRR abs/2103.08057 (2021).

[108] M. Zhou, J. Zhang, G. Adomavicius, Longitudinal impact of preference biases on recommender systems' performance, Available at SSRN, https://dx.doi.org/10.2139/ssrn.3799525 (2021).

[109] G. Adomavicius, D. Jannach, S. Leitner, J. Zhang, Understanding longitudinal dynamics of recommender systems with agent-based modeling and simulation, in: SimuRec Workshop at ACM RecSys 2021, 2021.

[110] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, C. Goodrow, Fairness in recommendation ranking through pairwise comparisons, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, 2019, pp. 2212–2220.

[111] M. J. Kusner, J. Loftus, C. Russell, R. Silva, Counterfactual fairness, in: Advances in Neural Information Processing Systems, Vol. 30, 2017.

[112] Y. Li, H. Chen, S. Xu, Y. Ge, Y. Zhang, Towards personalized fairness based on causal notion, in: 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, 2021, pp. 1054–1063.

[113] G. Cornacchia, F. Narducci, A. Ragone, A general model for fair and explainable recommendation in the loan domain (short paper), in: Joint Workshop Proceedings of the 3rd Edition of Knowledge-aware and Conversational Recommender Systems (KaRS) and the 5th Edition of Recommendation in Complex Environments (ComplexRec) co-located with 15th ACM Conference on Recommender Systems (RecSys 2021), Vol. 2960 of CEUR Workshop Proceedings, 2021.

[114] S. Seymen, H. Abdollahpouri, E. C. Malthouse, A unified optimization toolbox for solving popularity bias, fairness, and diversity in recommender systems, in: Proceedings of the 1st Workshop on Multi-Objective Recommender Systems (MORS 2021) co-located with 15th ACM Conference on Recommender Systems (RecSys 2021), Vol. 2959 of CEUR Workshop Proceedings, 2021.

[115] H. Yadav, Z. Du, T. Joachims, Policy-gradient training of fair and unbiased ranking functions, in: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1044–1053.

[116] F. M. Harper, J. A. Konstan, The MovieLens Datasets: History and Context, ACM Trans. Interact. Intell. Syst. 5 (4) (2015).

[117] J. Misztal-Radecka, B. Indurkhya, Bias-aware hierarchical clustering for detecting the discriminated groups of users in recommendation systems, Information Processing & Management 58 (3) (2021) 102519.

[118] S. Yao, B. Huang, Beyond parity: Fairness objectives for collaborative filtering, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17, 2017, p. 2925–2934.

[119] M. Stratigi, H. Kondylakis, K. Stefanidis, Fairness in group recommendations in the health domain, in: 33rd IEEE International Conference on Data Engineering, ICDE 2017, San Diego, CA, USA, April 19-22, 2017, IEEE Computer Society, 2017, pp. 1481–1488.

[120] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, J. Vertesi, Fairness and abstraction in sociotechnical systems, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, 2019, p. 59–68.

[121] D. Jannach, M. Zanker, A. Felfernig, G. Friedrich, Recommender Systems - An Introduction, Cambridge University Press, 2010.

[122] D. Jannach, M. Zanker, M. Ge, M. Gröning, Recommender systems in computer science and information systems - a landscape of research, in: 13th International Conference on Electronic Commerce and Web Technologies (EC-Web 2012), 2012, pp. 76–87.

[123] D. Serbos, S. Qi, N. Mamoulis, E. Pitoura, P. Tsaparas, Fairness in package-to-group recommendations, in: Proceedings of the 26th International Conference on World Wide Web, WWW 2017, 2017, pp. 371–379.

[124] G. Adomavicius, A. Tuzhilin, Context-aware recommender systems, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, 2015, pp. 191–226.

[125] I. Koutsopoulos, M. Halkidi, Efficient and fair item coverage in recommender systems, in: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), IEEE, 2018, pp. 912–918.

[126] H. Abdollahpouri, R. Burke, B. Mobasher, Managing popularity bias in recommender systems with personalized re-ranking, in: Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, 2019, pp. 413–418.

[127] Z. Zhu, J. Wang, J. Caverlee, Measuring and mitigating item under-recommendation bias in personalized ranking systems, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, p. 449–458.

[128] V. W. Anelli, A. Bellogín, T. Di Noia, D. Jannach, C. Pomo, Top-n recommendation algorithms: A quest for the state-of-the-art, in: UMAP '22, 2022.

[129] T. Giannakas, P. Sermpezis, A. Giovanidis, T. Spyropoulos, G. Arvanitakis, Fairness in network-friendly recommendations, in: 22nd IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, WoWMoM 2021, 2021, pp. 71–80.

[130] E. Yalcin, A. Bilge, Investigating and counteracting popularity bias in group recommendations, Inf. Process. Manag. 58 (5) (2021) 102608.

[131] M. Stratigi, J. Nummenmaa, E. Pitoura, K. Stefanidis, Fair sequential group recommendations, in: Proceedings of the 35th Annual ACM Symposium on Applied Computing, 2020, pp. 1443–1452.

[132] H. Oosterhuis, Computationally efficient optimization of plackett-luce ranking models for relevance and fairness, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1023–1032.

[133] A. Delic, J. Neidhardt, T. N. Nguyen, F. Ricci, An observational user study for group recommender systems in the tourism domain, J. Inf. Technol. Tour. 19 (1-4) (2018) 87–116.

[134] O. E. Gundersen, S. Kjensmo, State of the art: Reproducibility in artificial intelligence, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), 2018, pp. 1644–1651.

[135] P. Cremonesi, D. Jannach, Progress in recommender systems research: Crisis? what crisis?, AI Magazine 42 (3) (2021) 43–54.

[136] A. Bellogín, A. Said, Improving accountability in recommender systems research through reproducibility, User Model. User Adapt. Interact. 31 (5) (2021) 941–977.

[137] A. Narayanan, 21 definitions of fairness and their politics, Tutorial at FAT* 2018 (2018).

[138] F. Silveira, C. Diot, URCA: pulling out anomalies by their root causes, in: INFOCOM 2010. 29th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 2010, pp. 722–730.

[139] R. Mehrotra, J. McInerney, H. Bouchard, M. Lalmas, F. Diaz, 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, CIKM 2018, 2018, pp. 2243–2251.

[140] I. Koprinska, K. Yacef, People-to-people reciprocal recommenders, in: Recommender Systems Handbook, Springer, 2015, pp. 545–567.

[141] A. S. Koshiyama, E. Kazim, P. C. Treleaven, Algorithm auditing: Managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms, Computer 55 (4) (2022) 40–50.

[142] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.