



Fairness in recommender systems: research landscape and future directions

Yashar Deldjoo¹ · Dietmar Jannach² · Alejandro Bellogin³ ·
Alessandro Difonzo¹ · Dario Zanzonelli¹

Received: 27 June 2022 / Accepted in revised form: 17 March 2023
© The Author(s) 2023

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. Afterward, through a review of more than 160 scholarly publications, we present an overview of how research in this field is currently operationalized, e.g., in terms of general research methodology, fairness measures, 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 and questions of the underlying normative claims and what represents a fair recommendation in the context of a given application are often not discussed in depth. These observations call for more interdisciplinary research to address fairness in recommendation in a more comprehensive and impactful manner.

Keywords Recommender systems · Fairness · Survey

✉ Yashar Deldjoo
deldjooy@acm.org

Dietmar Jannach
dietmar.jannach@aau.at

Alejandro Bellogin
alejandro.bellogin@uam.es

- ¹ Polytechnic University of Bari, Bari, Italy
- ² University of Klagenfurt, Klagenfurt, Austria
- ³ University Autonomous of Madrid, Madrid, Spain

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 (Jannach and Jugovac 2019). Only in the most recent years was more awareness 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 (Ntoutsis et al. 2020; Trattner et al. 2022). One important ethical question in this context is that of the *fairness* of a recommender system, see (Burke 2017; Ekstrand et al. 2022), reflecting related discussions on the more general level of *fair machine learning* and *fair AI* (Mehrabian et al. 2021; Barocas et al. 2019; Ntoutsis et al. 2020).

During the last few years, researchers have discussed and analyzed different dimensions in which a recommender system should be fair or vice versa.

Given the nature of fairness as a social construct, it, however, seems difficult or even impossible (Ekstrand et al. 2022), to establish a general definition of what represents a fair recommendation. In addition to the subjectivity of fairness, there are frequently competing stakeholder interests to account for in real-world recommendation contexts (Naghiaei et al. 2022; Abdollahpouri et al. 2020a).

With this survey, we aim to provide an overview of what has been achieved in this emerging area so far and highlight potential research gaps. Specifically, drawing on an analysis of more than 150 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. Selbst et al. (2019), calling for more faceted and multi-disciplinary approaches to research in fairness-aware recommendation.

The paper is organized as follows. Next, in Sect. 2, we lay out the motivation behind this survey in more detail, and we present the essential notions used to characterize fairness in the literature. Section 3 then presents our methodology to identify and categorize relevant research works. Section 4 represents the main part of our study, which paints the current research landscape of fairness in recommender systems in various dimensions, e.g., in terms of the addressed fairness problem and the chosen research methodology. In Sect. 5, we then reflect on these observations and identify open challenges and possible future research directions.

2 Background and foundations

2.1 Examples of unfair recommendations

In the general literature on Fair ML/AI, an important application case is the automated prediction of recidivism by convicted criminal. 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 (Angwin et al. 2016). In the context of our present work, such use cases of ML-based decision-support systems are not in 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, several situations also in typical application scenarios of RS where many people might think 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.

Nonetheless, questions of fairness in RS extend beyond the consumer's perspective. In reality, a recommendation service often involves multiple stakeholders (Abdollahpouri et al. 2020a). On music streaming platforms, for example, we have not only 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 in increasing their visibility through recommendations. On the other hand, platform providers might seek to maximize 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 being

exposed to consumers, which might be considered *unfair to providers*. Finally, there are also use cases where recommendations may have *societal* impact, particularly on news and social media sites. Some may 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 against certain user groups.

As we will see later, different notions of fairness exist in the literature. What is important, however, is that in any discussed scenario, there are certain ethical questions or principles which are put at stake, and these are usually related to some underlying normative claims (Srivastava et al. 2019; Cooper 2020). Our research, however, indicates that these normative claims are often not unpacked and discussed to a sufficient extent in today's research on fairness in recommender systems. For instance, it may be argued that the issue with an e-commerce site optimizing for profit is not that it does so, but rather that it does so while misleading people into believing that recommendations are tailored to their needs. In situations such as this, the distinction between unfair and deceptive business activities can easily get blurred.

We note here that being fair to consumers or society in the bespoke examples may, in turn, also service 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 inappropriately into bad positions. The current draft of the European Commission's Digital Service Act¹ can be seen as a prime example where recommender systems and their potential harms are explicitly addressed in legal regulations, as it "*calls for more fairness, transparency and accountability for digital services' content moderation processes, ensuring that fundamental rights are respected, and guaranteeing independent recourse to judicial redress.*"

Overall, several 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 behavior that may be considered unfair. For example, in Ekstrand et al. (2022), the authors report that unfairness can arise in many places, either in society, in the observations that form our data, and in the construction, evaluation, and application of decision support models. Similarly, in Ashokan and Haas (2021), the authors classify the biases in a computing system as pre-existing bias, technical bias, and emergent bias, whereas in Olteanu et al. (2019) the authors differentiate between issues introduced when collecting social data (in general, not focused on recommender systems), introduced while processing such data, pitfalls that occurred when analyzing data, and issues with the evaluation and interpretation of the findings. Herein, our discussions are based on

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

insights from these and other earlier works, aiming to summarize and highlight the main causes of unfairness reported in the literature.

One first common issue mentioned in the literature is that the data on which the machine learning model is trained is biased (Chen et al. 2022; Olteanu et al. 2019). Such biases might, for example, result from 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 reinforces existing biases or skewed distributions in the underlying data. Differences between recommendation algorithms in terms of reinforcing popularity biases and concentration effects were, for example, examined in Jannach et al. (2015). In some cases, the machine learning model might also directly consider a “protected characteristic” (or a proxy thereof) in its predictions (Ekstrand et al. 2022). 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 (fairness through unawareness) (Grgic-Hlaca et al. 2016). Despite its appealing simplicity, this definition has a clear issue, as sensitive characteristics may have historically affected non-sensitive characteristics (e.g., a person’s GPA may have been influenced by their socioeconomic status). In order to adjust for biases in data collection or historical outcomes, it has been argued that, in fact, protected characteristics must be taken into account to place other observable features in context (Kusner et al. 2017).

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 simultaneously keeping users satisfied as much as possible (Ghanem et al. 2022; Jannach and Adomavicius 2017). 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” (Ekstrand et al. 2022), 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 even though the real reason certain characteristics are regarded protected is because of historical discrimination or subordination, where redress is necessary. Rather, the question is to what extent this is reflected in the data or how this unfairness influences the fairness goals.

In general, the underlying reasons also determine *where* in a machine learning pipeline² interventions can or should be made to ensure fairness (or to mitigate unfairness). In a common categorization, (Mehrabi et al. 2021; Shrestha and Yang 2019; Pitoura et al. 2022; Zehlike et al. 2022a), 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*

² Consider Ashokan and Haas (2021), where the authors show that biases may occur in a typical machine learning pipeline from data generation, over the model building and evaluation, to deployment and user interaction.

(i.e., mathematically expressed) goal. In the 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 Sect. 4, such a goal is often formalized as a target distribution and/or as an evaluation metric to gauge the level of existing or mitigated fairness.

2.3 Notions of fairness

When dealing with phenomena of unfairness such as those outlined, and when our purpose is to prevent or mitigate such phenomena, a question arises: what do we consider fair in general and in a particular application context? Fairness, in general, is fundamentally a societal construct or a *human value*, which has been discussed for centuries in 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 and Rubin 2018; Verma et al. 2020) for a high-level discussion of the definition of fairness in machine learning and ranking algorithms, or Mulligan et al. (2019) for the relationship to social science conception of fairness. As we will see in the remainder of this survey, fairness is a complex concept with multiple perspectives. Consequently, there are numerous definitions, but none of them appear to be exhaustive.

In general, the societal constructs around fairness mainly depend on how moral standards or dilemmas are addressed: either through *descriptive* or *normative* approaches (Srivastava et al. 2019). While normative ethics involves creating or evaluating moral standards to decide what people should do or whether their current moral behavior is reasonable, descriptive (or comparative) ethics is a form of empirical research into the attitudes of individuals or groups of people towards morality and moral decision-making. As mentioned above, normative claims are often not explicitly specified in existing research, both in general machine learning and in recommender systems research. In fact, it was already recommended in earlier research to make these assumptions more explicit (Cooper 2020). From our study of the literature, we observe that a majority of the works did not clarify what the actual normative claim is being addressed or who is representing or making such claims.

As a possible consequence of this problem, we also observe that researchers, in most cases, do not refer to a specific public discussion of the topic at hand. For many papers on recommender systems, there is, for example, no indication or evidence that there is a public debate outside computer science, e.g., whether or not it is fair to recommend niche movies. Nonetheless, it is true that there actually are areas, like job recommendation, where a public discussion takes place, e.g., about discrimination and what normative claims are agreed to be addressed.

The primary notions of fairness that will be used throughout this review—as extracted from the aforementioned literature and recent surveys (Li et al. 2022; Wang et al. 2022b)—are presented next and further expanded in Section 4.6. We emphasize that these definitions present a specific perspective on defining the concept of fairness. They are, however, not necessarily *orthogonal* and *all-encompassing*. Table 1 shows

examples of fictitious statements of a user regarding unfairness in a job recommendation scenario under different notions of fairness.

- *Group vs. individual*: 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. Group fairness, in contrast, aims to ensure that “different groups have similar experience” (Ekstrand et al. 2022), i.e., protected groups receive similar benefits from the decision-making as others. Typical groups in such a context are a majority or dominant group and a protected group (e.g., an ethnic minority). Since this may be too simplistic, other authors state *we are all equal* as the fundamental logic underlying group fairness (Friedler et al. 2021), asserting their equivalence as a starting point.
- *Process vs. outcome*: Process (or: treatment) unfairness means that individuals with similar non-sensitive attributes receive different outcomes solely due to the difference in sensitive features. Outcome (or: impact) unfairness occurs when a system produces outputs that benefit (harm) a group of individuals sharing a sensitive attribute value more frequently than other groups (Zafar et al. 2017). Put it differently, process fairness assesses aspects such as the data used, the decision-making principles of the system, and the causal association between inputs and outputs. In contrast, outcome fairness disregards the internal operation of the system and concentrates solely on the equitable distribution of rewards (Amigó et al. 2023).
- *Direct vs. indirect*: Fairness can also be analyzed based on whether particular sensitive feature holders are directly harmed or not (Council et al. 2004). Direct fairness refers to situations in which persons receive less favorable treatment based on protected characteristics such as race, religion, or gender. When the reasons for the discrimination are only tenuously connected to (or identical to) the protected characteristic, we have indirect fairness.³ For example, some institutions use the location of candidates as a *proxy* for an overtly discriminating characteristic (e.g., race) (Zhang and Bareinboim 2018).
- *Statistical vs. predictive parity*: In machine learning, fairness definitions fundamentally seek some sort of equity on various portions of the *confusion matrix* used for binary classification evaluation. Statistical parity is independent of the actual value and requires protected group members to have an equal positive prediction rate. Predictive parity employs the actual outcome and requires that the model’s precision (or accuracy) is comparable for all subgroups under consideration.
- *Static vs. dynamic*: In static fairness, the recommendation environment is fixed during the recommendation process; hence, the user activity level is assumed to remain unchanged. Dynamic fairness definitions, on the other hand, integrate the (typical) dynamic attribute of most recommender systems, which needs to consider new user interactions, new items, or continually evolving user groups.
- *Associative vs. causal*: Associative fairness metrics are computed based on data and do not allow reason about the causal relations between the features and the

³ The term *redlining* (Corbett-Davies and Goel 2018) is analogous to the concept of indirect unfairness wherein a non-sensitive characteristic (such as geography) is used as a proxy for a more personal quality (such as race or socioeconomic status).

Table 1 Examples for possible statements around different notions of fairness in the context of a recommender system for jobs

<i>Group</i>	vs. <i>Individual</i>
Compared to men, women are recommended low-paying occupations!	My friend Elisa and I had similar GPA, qualifications and skills, but she got better job recommendations!
<i>Process</i>	vs. <i>Outcome</i>
My friend John and I had similar GPA, qualifications, and skills, but he got better suggestions only because he's a man!	Relevant higher-paying jobs get recommended to white people rather than black!
<i>Direct</i>	vs. <i>Indirect</i>
I am receiving worse recommendations only because of my skin color!	People from south Italy receive worse job recommendations by the system!
<i>Statistical parity</i>	vs. <i>Predictive parity</i>
My group should receive as many good recommendations as other groups!	Among people who are recommended for the job, there is a smaller share of qualified people from my group than from other groups!
<i>Static</i>	vs. <i>Dynamic</i>
The system achieved to be fair just once, in a different job market, but now employees' goals and priorities have changed!	The system accounts for shifts in our tastes and needs, and can prefer me today if it preferred someone else yesterday!
<i>Associative</i>	vs. <i>Causal</i>
If you are black-skinned, you are historically more likely to be discriminated against!	Had I not been black-skinned, would I have received that recommendation?

decisions. Causal fairness definitions, on the other hand, are usually defined in terms of (non-observable) interventions and counterfactuals and tend to consider the additional structural knowledge of the system regarding how variables propagate on a causal model (Li et al. 2022).

Other categorizations can be found in the literature, based on *short-term vs. long-term* considerations (according to the duration of the fairness requirements), *granularity* (whether the system applies the same fairness notion to everyone or if users could decide how they want to be treated by the system), *transparency* (to discriminate notions that are explainable from those that are a black box), or *depending on the associated fairness concept* (such as consistent, calibrated, counterfactual, Rawlsian maximin, envy-free, and maximin-shared) (Li et al. 2022; Wang et al. 2022b; Amigó et al. 2023). An in-depth discussion of these—sometimes even incompatible (Verma and Rubin 2018; Amigó et al. 2023)—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 Sect. 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 Selbst et al. (2019).

2.4 Related concepts: responsible recommendation and biases

Issues of fairness are often discussed within the broader area of *responsible* recommendation (Elahi et al. 2022; Ekstrand et al. 2022; Di Noia et al. 2022), with the key dimensions *generalizability*, *robustness* (Deldjoo et al. 2020, 2021c), *privacy* (Anelli et al. 2021; Friedman et al. 2015), *interpretability* (Tintarev and Masthoff 2022; Deldjoo et al. 2023), and *fairness*, with the definitions of these concepts blurring as we progress through the list. In Elahi et al. (2022), the authors, in particular, discuss the 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 setting, for example, when one gender or race is treated differently just based on this attribute, as when suggesting images for a specific profession. Fairness is therefore seen as a particular aspect of responsible recommendation in Elahi et al. (2022). A similar view is taken in Ekstrand et al. (2022), 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* (Chen et al. 2022; Baeza-Yates 2018). 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 in part be the result of an already existing recommender system and, hence, biased by what is shown to users. 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 Ekstrand et al. (2022). However, the use of the term is inconsistent in the literature, as also observed in our work in Chen et al. (2022) and in our work. In some early papers, *bias* is used almost synonymously with *fairness*. In Friedman and Nissenbaum (1996), 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. Instead, it all depends on the underlying ethical principles and normative claims, as discussed before. Moreover, an in-depth discussion and systematic comparison of

various forms of biases is beyond the scope of our work; we instead refer the reader to Chen et al. (2022), where different forms of biases are discussed in more depth.

3 Research methodology

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

3.1 Paper collection process

We adopted a mixed and semi-systematic approach to identify relevant research papers.⁴ In the first step, we identified relevant research papers by querying the DBLP⁵ digital library with predefined search terms and a set of explicit criteria for inclusion and exclusion. Afterwards, to include relevant papers which did not match the search terms in this still-evolving field, we (a) applied a snow-balling procedure and (b) relied on researcher experience to identify other relevant papers that were published in focused outlets.

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 in its 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 only as preprints on arXiv.org⁶ and we removed survey papers. We then manually scanned the remaining 268 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.
- 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 157 papers. The papers were read by at least two researchers and categorized in various dimensions, see Sect. 4.

⁴ We note here that our work is not intended to be a systematic literature review in the strict sense of Kitchenham et al. (2009), but rather aims to outline a broader picture of current research activities.

⁵ <https://dblp.org/>.

⁶ Note that DBLP indexes arXiv papers.

3.2 Relation to previous surveys

A number of related surveys were published in the last few years. The survey provided by Chen et al. (2022) focuses on biases in recommender systems, and connects different types of biases, e.g., popularity biases, with questions of fairness, see also (Abdollahpouri et al. 2020b). Note that bias mitigation in recommendation mostly focuses on increasing the accuracy or robustness of the recommendations through debiasing approaches, rather than on promoting fairness.

The recent monograph by Ekstrand et al. (2022) discusses fairness aspects in the broader context of *information access* systems, an area that 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. This study differs from our work in that our objective is not to give a fresh classification of fairness concepts and methods found in the literature. Instead, our main objective 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. Ekstrand et al., on the other hand, focus more generally on future directions in this area.

Different survey papers were published also in the more general area of fair machine learning or fair AI, as mentioned above (Mehrabi et al. 2021; Barocas et al. 2019). 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.

Very recently, while we conducted our research, a number of alternative surveys on fairness in recommender systems have become available as preprints or peer-reviewed publications, including Pitoura et al. (2022), Zehlike et al. (2022b), Wang et al. (2022b), and Li et al. (2022). Clearly, there is a certain overlap of our survey and these recent publications, e.g., in terms of the used taxonomy of fairness-related aspects. Note, however, that unlike some of these papers, e.g., Li et al. (2022), Pitoura et al. (2022), our aim is *not* to establish a new taxonomy or to discuss the technical details of specific computational metrics or algorithmic approaches that were proposed in the past literature. Instead, our aim is to paint a landscape of existing research and to thereby identify potential research gaps. In that context, our work has similarities with the work by Wang et al. (2022b), who reviewed and categorized 60 recent works on fairness in recommender systems. While our survey involves a larger number of papers, Wang et al. dive deeper into the technicalities of particular approaches, which is not the focus of our work. Here, in contrast, we aim to paint a broader picture of today's research activities and existing gaps without entering into the technical specifics of existing approaches. Moreover, our work also emphasizes more on evaluation aspects and on potential methodological issues in this research area. The recent work by Zehlike et al. (2022b), finally, mainly discusses individual research works in detail, also including more general ones on learning-to-rank. The overlap with this work, except for the discussion of different dimensions of fairness, is therefore limited.

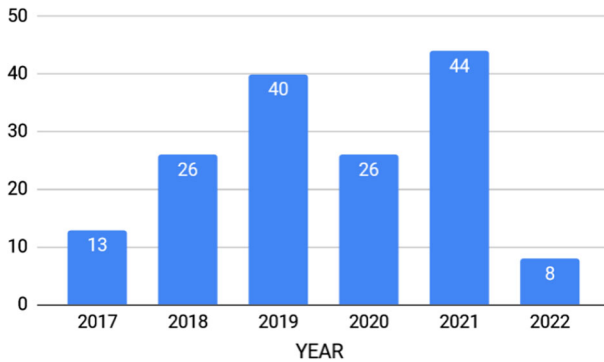


Fig. 1 Number of papers published per year. The entire number of papers sum up to 157

In general, the goal of these existing works is mainly to review and synthesize the various existing approaches so far to design fair recommender systems and to evaluate them. The goal of our work is indeed different, as we aim to analyze and quantify which notions of fairness the research community is working on and how the research problem is operationalized. Differently from previous surveys, our study can therefore inform about the less frequently studied areas, and thus potential gaps, of fairness research in a quantitative manner. Moreover, our analyses of the applied research methodologies reveal a very strong predominance of data-based experiments, which rely on abstract computational metrics and do not involve humans in the loop. We, therefore, believe that our survey complements existing surveys well.

4 Landscape of fairness research in recommender systems

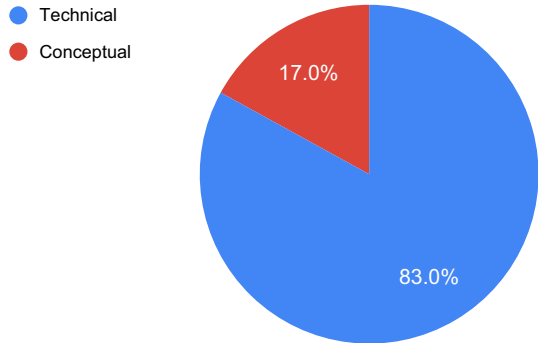
In this section, we categorize the identified literature along different dimensions to paint a landscape of current research and to identify existing research gaps.

4.1 Publication activity per year

Interest in fairness in recommender systems has been constantly growing over the past few years. Figure 1 shows the number of papers per year that were considered in our survey. Questions of fairness in information retrieval have been discussed for many years, see, e.g., Pedreshi et al. (2008) for an earlier work. The area has been consistently growing since then, leading also to the establishment of dedicated conference series like the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT).⁷ 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 published as late as in 2017.

⁷ A number of related events have been recently connected through the ACM FAccT Network, <https://facctconference.org/network/>.

Fig. 2 Technical vs. Conceptual Papers



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 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 (Jannach et al. 2016, 2021). 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 based on the research contributions of the surveyed papers 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 (Schelenz 2021), exploratory studies that investigate user perceptions of fairness (Sonboli et al. 2021), or discussions about how difficult it is to audit these types of systems (Krafft et al. 2020).

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 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 some pre-existing and mathematically defined optimization goals are 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 fairness goals and (b) in finding or designing operationalizable optimization goals (e.g., a computational metric) which represent reliable measures or proxies for the given fairness goals.

4.3 Categorization of notions of fairness in literature

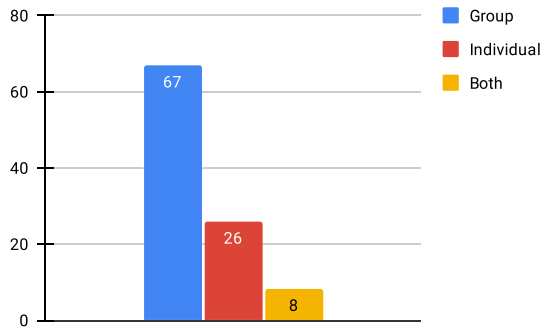
In Li et al. (2021c), a taxonomy of different notions of fairness was introduced: group vs. individual, single-sided vs. multi-sided, static vs. dynamic, and associative vs. causal fairness; see also our discussions in Sect. 2.3. In the following, we review the literature following this taxonomy.⁸

Group vs. individual fairness 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 (Binns 2020). 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 (Dwork et al. 2012). Ideas of individual fairness in machine learning were discussed in an early work in Dwork et al. (2012), 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 (Binns 2020).

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

⁸ Each paper was categorized by at least two researchers, and potential discrepancies were resolved through a discussion process. The same process was applied to categorize the papers also in other dimensions as discussed later in this section.

Fig. 3 Group vs. Individual Fairness



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 other 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 internally⁹ from the interaction data (e.g., based on user activity level, mainstreamness, or item popularity) (Abdollahpouri et al. 2021; Li et al. 2021a). In Table 2, 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) (Burke 2017) or multi-sided fairness..

Additionally, it is possible to observe in RS settings that these sensitive attributes may be provided by external providers as demographic metadata (for example, user's gender, age, occupation), or they may be extracted from user-item interaction data, for example, dividing users based on their level of activity (i.e., active vs. inactive users), or the types of items they consume (e.g., mainstream-users vs. non-mainstream). Here a related concept is *obfuscation* (Slokom et al. 2021), which is a strategy for privacy protection to conceal sensitive information. Fairness and privacy can be considered as interwoven under obfuscation, as described by Dwork et al. (2012) and Pessach and Shmueli (2022), where a violation of privacy can lead to unfairness due to an adversary's capacity to infer sensitive information about an individual and utilize it in a discriminatory manner.

⁹ We should note that we found no example where the reliability of these implicitly computed attributes was analyzed. Usually, authors use explicit thresholds to assign users/items to groups (Li et al. 2021a; Xiao et al. 2020) or percentiles from distributions based on a variable of interest, such as item popularity (Abdollahpouri et al. 2021; Deldjoo et al. 2021a).

Table 2 Overview of common attributes used when addressing fairness concepts from the perspectives of consumers, providers, or both

Attribute	
Goal 1: Consumer Fairness	
Target: <i>Demographic parity</i> – sensitive attributes are attained by birth and not under a user’s control.	<ul style="list-style-type: none"> – Gender (Deldjoo et al. 2021a, b; Wu et al. 2021a; Gorantla et al. 2021; Wang et al. 2021; Ghosh et al. 2021a; Wan et al. 2020; Edizel et al. 2020; Tsintzou et al. 2019; Mansoury et al. 2019; Lin et al. 2019; Geyik et al. 2019; Xia et al. 2019; Burke et al. 2018; Chakraborty et al. 2017; Farnadi et al. 2018; Burke et al. 2017; Riederer and Chaintreau 2017) – Race (Gorantla et al. 2021; Ghosh et al. 2021a; Zheng et al. 2018; Zhu et al. 2018b; Chakraborty et al. 2017; Zhu et al. 2018c; Riederer and Chaintreau 2017) – Age (Deldjoo et al. 2021a; Bobadilla et al. 2021; Sühr et al. 2021; Melchiorre et al. 2021; Gorantla et al. 2021; Farnadi et al. 2018) – Nationality (Weydemann et al. 2019) and Location (Riederer and Chaintreau 2017) – Occupation (Farnadi et al. 2018)
Target: <i>Merit-based fairness</i> – attained through a user’s merit over time	<ul style="list-style-type: none"> – Education (Sühr et al. 2021; Gómez et al. 2021) – Income (Sühr et al. 2021)
Target: <i>Behavior-oriented fairness</i> – attained based on a user’s engagement with the system/item catalog.	<ul style="list-style-type: none"> – User (in)activeness (Hao et al. 2021; Li et al. 2021a; Xiao et al. 2020; Fu et al. 2020; Chakraborty et al. 2019) – User (non)mainstreaminess (Abdollahpouri et al. 2020b, 2021)
Target: <i>Other emerging attributes</i>	<ul style="list-style-type: none"> – Physio/psychological (Wan et al. 2020; Htun et al. 2021) – Sentiment-based (Lin et al. 2021)
Goal 2: Provider Fairness	
Target: <i>Item producer/creator</i> – sensitive attribute based on who the item producer is	<ul style="list-style-type: none"> – News author (Gharahighehi et al. 2021), music artist (Ferraro 2019), movie director (Boratto et al. 2021b)
Target <i>Producer’s demographic or general information</i> – sensitive attribute based on to which demographic group the item producer belongs, e.g., male vs. female artists.	<ul style="list-style-type: none"> – Gender (Kirnap et al. 2021; Boratto et al. 2021b; Shakespeare et al. 2020; Xia et al. 2019), geographical region (Gómez et al. 2021)
Target: <i>Item information</i> – sensitive attribute based on the item information itself	<ul style="list-style-type: none"> – Price and brand (Deldjoo et al. 2021a; Dash et al. 2021), geographical region (Liu et al. 2020; Burke et al. 2018)
Target: <i>Interaction-oriented fairness</i> – sensitive attribute based on the interactions observed on items e.g., popularity	<ul style="list-style-type: none"> – Popularity (Deldjoo et al. 2021a; Dong et al. 2021; da Silva et al. 2021; Wundervald 2021; Borges and Stefanidis 2021; Ge et al. 2021; Sun et al. 2019; Weydemann et al. 2019; Abdollahpouri et al. 2019b; Zhu et al. 2018a), cold items (Zhu et al. 2021)

Table 2 continued

	Attribute
Target: <i>Other emerging attributes</i>	– Premium membership (Deldjoo et al. 2019), sentiment and reputation (Lin et al. 2021; Zhu et al. 2020a)
Target: <i>Non-sensitive attributes</i>	– Movie and music genre (Tsintzou et al. 2019; Lin et al. 2019; Rastegarpanah et al. 2019; Ferraro 2019)
Goal 3: Consumer Provider Fairness (Multi-sided Fairness)	
Target: Combinations of two targets from C-Fairness and P-Fairness	– Same category of sensitive attributes for both users and items (e.g. <i>behavior-oriented</i>) (Naghiaei et al. 2022; Rahmani et al. 2022a, b; Lin et al. 2021; Abdollahpouri et al. 2019b; Burke et al. 2018) – Different categories of sensitive attributes (Deldjoo et al. 2021a, 2019; Mansoury et al. 2019; Tsintzou et al. 2019; Weydemann et al. 2019; Xia et al. 2019; Rahmani et al. 2022c)

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 (Gómez et al. 2021). In these cases, user demographics may play a major role, together with other factors such as popularity, expertise, and availability at a certain point in time. 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 Li et al. (2021a) 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 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 per se. In fact, merely observing different levels of recommendation accuracy for more active and less active users may not be enough to conclude that a system is unfair. Instead, it is important to carefully elaborate on the underlying reasons and the related normative claims. Some particular user groups may for example have had fewer opportunities to engage with a system.

Recent studies have also focused on two-sided CP-Fairness, as illustrated in Naghiaei et al. (2022); Rahmani et al. (2022b). 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 Naghiaei et al. (2022), 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

providers (suppliers), see also (Li et al. 2021c), which differentiates between user and item fairness. In these research works, the idea often is to avoid an unequal (or: unfair) *exposure* of items from different providers, e.g., artists in a music recommendation scenario. The term *item fairness*, although used in the literature, may however not be optimal. In reality, it might be argued that this perspective is only important because the item providers—hence, other people or organizations—are actually impacted and, therefore, the underlying fairness concept aims to convey some sense of social justice related to people.

In some works, e.g., Boratto et al. (2021a), the *popularity* of items is considered an important attribute. Typical goals in that context are to give fair exposure to items that belong to the long tail, or to include a combination of popular and less popular items in a user-calibrated fashion (Abdollahpouri et al. 2021). In other research works that focus on fair item exposure, e.g., in Gupta et al. (2021), groups are defined based on attributes that are in practice not protected in legal terms or based on some accepted normative claim, e.g., the price range of accommodation. 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 (Zhao et al. 2022).

Fairness for items at the *individual* level, in particular for cold-start items, is for example discussed in Zhu et al. (2021). In general, as shown in Fig. 3, works that consider aspects of individual fairness are less frequently investigated than group fairness scenarios. An even smaller number of works addresses both types of fairness.

The definition from classical fair ML settings—similar individuals should be treated similarly—cannot always be directly transferred to recommendation scenarios. In Edizel et al. (2020), 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 similarly¹⁰. 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 (Masthoff and Delic 2022; Felfernig et al. 2018), 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.¹¹ In more recent works on group recommendation, in contrast, fairness is explicitly mentioned, e.g., in Htun et al. (2021),

¹⁰ It should be noted that if decisions would be based on the protected gender attribute, it would not be individual fairness. In the discussed work, however, the goal is to treat individuals similarly which have similar attributes (and not considering the gender attribute). This then represents an approach towards individual fairness according to the definition.

¹¹ Even though there are some strategies that are not fair, e.g., dictatorship, where one decides for the group (Masthoff and Delic 2022).

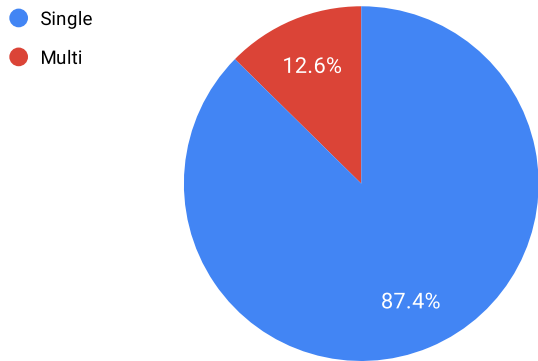
Kaya et al. (2020), Malecek and Peska (2021), maybe also due to the current interest in this topic. Notable works in this context are (Htun et al. 2021) and (Wang et al. 2022a), which are one of the few works in our survey which consider questions of fairness *perceptions*.

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 (Oh et al. 2011; Steck 2018; Jugovac et al. 2017). While this may not be inherent and directly related to individual or group fairness, this is the category from this section that better suits such an important (and popular) technique. In fact, from a conceptual point of view, one may see calibration as implementing a particular form of group fairness, without there being an explicitly protected attribute. In the entertainment domain, this might be the (implicit) group of independent movie lovers (Abdollahpouri et al. 2021); in the news domain, there may be a group of users who prefer a balanced information offering, e.g., in terms of political opinions. Applying calibration may then help to avoid that the independent movie lovers receive mainly recommendations of mainstream movies; and that vice versa independent movies obtain a higher chance of exposure.

More in general, calibration has been applied to either users—by considering age or gender as features to be calibrated against—or items—to compensate for popularity, but also to diversify with respect to item attributes such as genre (Bobadilla et al. 2021; Abdollahpouri et al. 2021; da Silva et al. 2021). Besides, in works like (Abdollahpouri et al. 2020b), calibration is considered as a quality of the recommendations, and the authors measure whether different users or groups experience varying levels of (mis)calibration in their recommendations, since this may indicate an unfair treatment on those populations. Nonetheless, as stated in Lin et al. (2020), calibrated recommendations in some domains (such as news or microblogging) might contribute to political polarization in society, so this technique is generally applied to consumer taste domains, where focused, less-diverse recommendations might be valued by users. Like for other fairness approaches, however, there must be an underlying normative claim that is addressed. Without an underlying normative claim, calibrating recommendations may in some cases merely be a matter of improved personalization and, thus, recommendation quality.

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 (Xiao and Benbasat 2007) 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

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



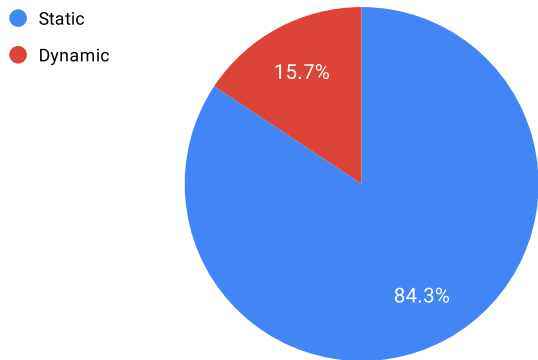
include consumers, service providers, suppliers of the recommendable items, or even society (Abdollahpouri et al. 2020a; Jannach and Bauer 2020).

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 (Jannach and Adomavicius 2017). In a similar vein, questions of fairness can be viewed from multiple stakeholders, leading to the concept of *multisided* fairness (Burke 2017), which might include the utility of system designer and other side-stakeholders in addition to the consumer and provider. As mentioned above, there can be fairness questions that are related to the providers of the items. Again, there can also be tradeoffs and in some ways incompatible notions of fairness, i.e., what may be a fair recommendation for users might be in some ways be seen to be unfair to item providers, e.g., when their items get limited exposure (Chaudhari et al. 2020).

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.

Among the few studies on multi-sided fairness, Abdollahpouri and Burke (2019) discusses techniques for CP-fairness in matching platforms such as Airbnb and Uber. In Rahmani et al. (2022a), the authors explore how adding contextual information such as geographical, temporal, social, and categorical affects the multi-aspect quality of POI suggestions, including accuracy, beyond-accuracy, fairness, and interpretability (see also Rahmani et al. 2022d for a discussion on a temporal bias). Patro et al. (2020) 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. In Anelli et al. (2023) the authors study the CP-Fairness in several graph CF models. Wu et al. (2021b) propose an individual-based perspective, where fairness is defined as the same exposure for all producers and the same NDCG for all consumers involved. Exposure in this work is defined based on the appearance of items of providers on top-n recommendation lists, where a higher ranking is assumed to lead to higher exposure.

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

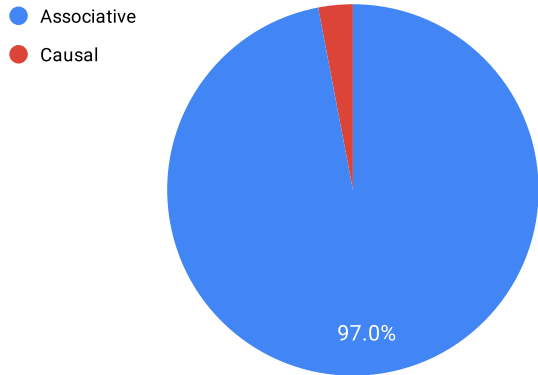


Static vs. dynamic fairness Another dimension of fairness research relates to the question whether the fairness assessment is done in a static or dynamic environment (Li et al. 2021c). In static settings, the assessment is done at a single point of time, as commonly done also in offline evaluations that 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 (Rohde et al. 2018; Mladenov et al. 2021; Ghanem et al. 2022; Zhou et al. 2021; Adomavicius et al. 2021) for simulation studies and related frameworks in recommender systems.

Figure 5 shows how many studies in our survey considered static and dynamic evaluation settings, respectively. Static evaluations are clearly predominant: we only found 16 works that consider dynamically changing environments. In Ge et al. (2021), 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 Liu et al. (2020), 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. Moreover, in Sonboli et al. (2020), a framework is proposed for the dynamic adaptation of recommendation fairness using *Social Choice*. The goal of this work is to arbitrate between different re-ranking methods, aiming to achieve a better accuracy-fairness tradeoff with respect to all sensitive features. On the other hand, works such as (Beutel et al. 2019) and (Deldjoo et al. 2021a) 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.

Fig. 6 Fairness Notions:
Associative vs. Causal Fairness



Associative vs. causal fairness The final categorization discussed in Li et al. (2021c) 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 (Kusner et al. 2017; Li et al. 2021b).

Figure 6 shows that there are only *three* works investigating recommendation fairness problems based on causality considerations. More specifically, in Cornacchia et al. (2021), the authors propose the use of counterfactual explanation to provide fair recommendations in the financial domain. An interesting alternative is presented in Li et al. (2021b), where the authors analyze the causal relations between the protected attributes and the obtained results. The third work we found in our review, Qiu et al. (2021), derives a causal graph to identify and analyze the visual bias of existing methods, so that spurious relationships between users and items can be removed.

One additional dimension we have discovered through our literature analysis is the use of *constraint-based approaches* to integrate or model fairness characteristics in recommender systems. In this context, these approaches may be seen as an alternative paradigm to associative and causal inference, which is based on explicit constraints and special techniques, often from multi-objective optimization, to achieve the desired fairness goals. For example, Hao et al. (2021) 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 (Seymen et al. 2021) or (Yadav et al. 2021)—define fairness as another constraint to be optimized by the algorithms. In Yadav et al. (2021), in particular, such a constraint is amortized fairness-of-exposure.

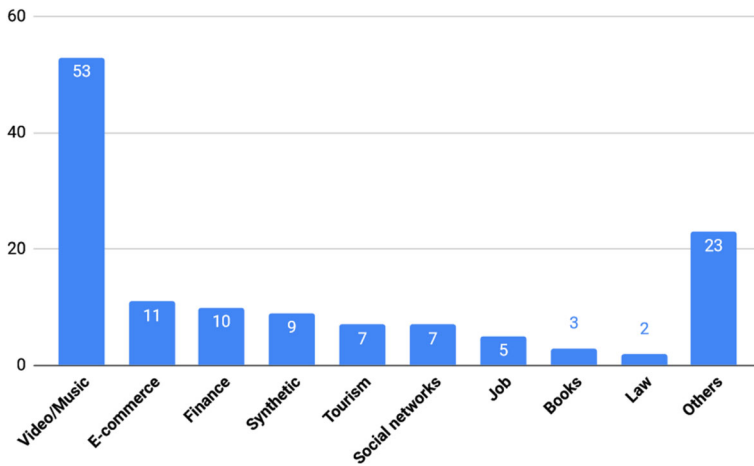


Fig. 7 Application domains of used datasets. Note that some studies rely on more than one dataset, and a number of theoretical or conceptual works do not provide experimental validation

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.¹² 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 Fig. 7.

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 widely adopted and probably overused case and have been used for all sorts of research in the past (Harper and Konstan 2015). 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

¹² The categorization of the papers was based on the datasets that were used for the empirical evaluations. We used higher-level categories of domains as done in earlier surveys, e.g., in Nunes and Jannach (2017), Jannach et al. (2012).

works aim to increase the exposure of long-tail items which are not too popular, see, e.g., Dong et al. (2021). In terms of *fairness towards users*, the technical proposal in da Silva et al. (2021) for example aims to serve users with recommendations that reflect their past diversity preferences with respect to movie genres. An approach towards *group fairness* is proposed in Misztal-Radecka and Indurkha (2021). 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.

In other domains beyond Video/Music (dominated, as mentioned above, by MovieLens datasets), fairness is characterized by the inherent properties of users and items in each particular domain. For example, in e-commerce the price or year of the item, or the helpfulness of the provided user's review are considered (Deldjoo et al. 2021a); in tourism, the user's gender and the business category are typically analyzed (Mansoury et al. 2019).

Continuing our discussions above, such notions of unfairness in the described application contexts 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. In another example, it might not be clear why recommending less popular items—which might in fact not be popular because of their limited quality—would make a system fairer, and equating bias (or skewed distributions) with unfairness in general seems questionable. Therefore, we iterate the importance of clearly specifying the underlying assumptions, hypothesis, and normative claims in any given research work on fairness. Otherwise it may remain unclear to what extent a particular system design or algorithmic approach will ensure or increase a system's level of fairness.

Similar questions arise when using calibration approaches to ensure fairness in a personalized, user-individual way. Considering, for example, a user fairness calibration approach like the one presented in da Silva et al. (2021), 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 recommendations of limited quality in terms of calibration for everyone is probably not one we would call unfair. However, note that there actually *may be* situations where calibration serve a certain fairness goal. Consider, for example, that a recommendation provider notices that users with niche tastes often receive item recommendations that are not interesting to them. This may happen when an algorithm too strongly focuses on mainstream items and when the used metrics do not reveal clearly that there are some user groups that are not served well. Under the assumption that users with niche tastes might also be users who are marginalized in other ways, e.g., when they are users who differ because of ethnicity or national origin, then improving calibration may indeed serve a fairness goal. These assumptions and claims however have to be made explicit, as otherwise it might just be an issue of whether the recommendation quality is measured in the right way.

In several cases, and independent of the particular application domain, it therefore seems that the addressed problem settings are not too realistic or remain artificial to a certain extent. 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 (Ge et al. 2021; Abdollahpouri et al. 2021; Yao and Huang 2017; Misztal-Radecka and Indurkha 2021; Hao et al. 2021; Tsintzou et al. 2019; Sun et al. 2019; Geyik et al. 2019; Stratigi et al. 2017). 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. Related phenomena of "abstraction traps" in fair ML were discussed earlier in Selbst et al. (2019). While abstraction certainly is central to computer science, the danger exists that central domain-specific or application-specific idiosyncrasies are abstracted away so that ML tools can be applied. In the end, the proposed solutions for the abstracted problem may then fail to properly account for the sometimes complex interactions between technical systems and the real world, and to respond to the "*fundamental tensions, uncertainties, and conflicts inherent in sociotechnical systems.*" (Selbst et al. 2019)

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) (Gunawardana et al. 2022; Jannach et al. 2010). 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 (Jannach et al. 2012), we find that offline experiments are the predominant form of research. Note that we here only consider 83 technical papers, and not the conceptual, theoretical, and analytic ones that we identified. Only in very few cases (6 papers), humans were involved in the experiments, and in even fewer cases (3 papers) we found reports of field tests. Regarding user studies, Htun et al. (2021) 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 Geyik et al. (2019), 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 Beutel et al. (2019), where the engagement with a large-scale recommender system in production is reported across sub-groups of users. We only found one paper that relied on interviews as a qualitative research method (Sonboli et al. 2021). Also, only very few papers used more than one experiment type, e.g., Serbos et al. (2017) where both a user study and an offline experiment were conducted.

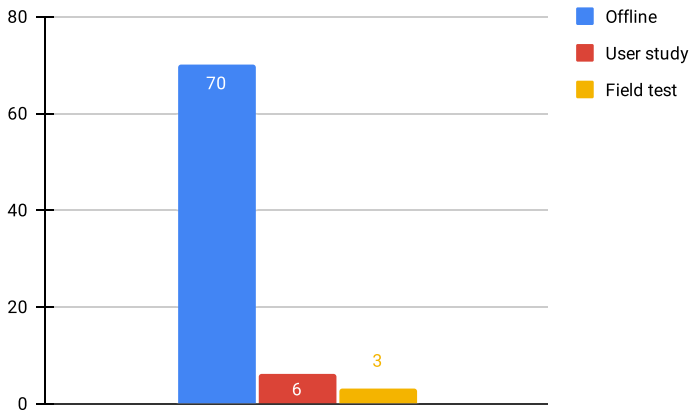


Fig. 8 Experiment types

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 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, as characterized above. In a recent work on job recommendations (Wang et al. 2022a), 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 (Adomavicius and Tuzhilin 2015), 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, see also (Li et al. 2022). 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.

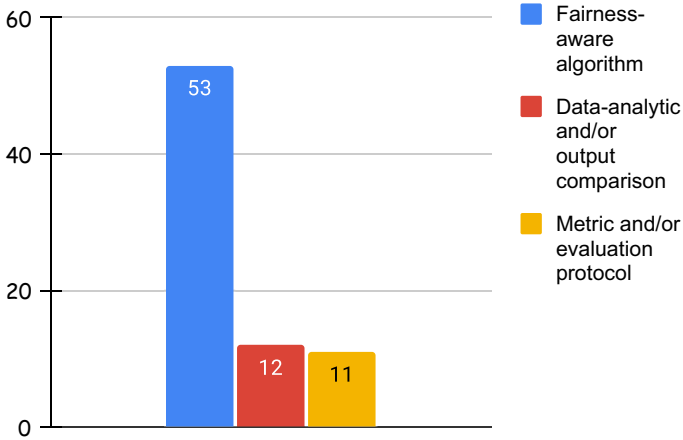


Fig. 9 Technical focus of papers

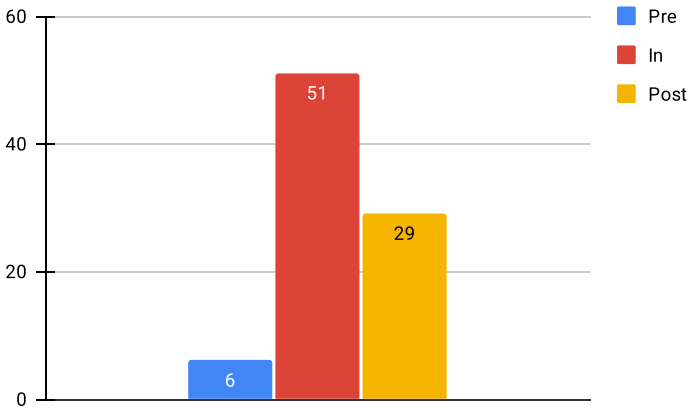


Fig. 10 Fairness step

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 particular domains. In the future, it might therefore be important to focus more on analytical research, as advocated also in Jannach and Bauer (2020), 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 work 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. In offline experiments, a variety of computational metrics are employed to evaluate the fairness of a set of recommendations. The choice of a certain fairness metric is mostly determined by the underlying concept of fairness, such as whether it is about individual or group fairness. In Table 3 and Table 4, we provide detailed lists of selected metrics used in the literature on fairness in recommender sys-

tems.¹³ We primarily organize the metrics along the common categorization of group fairness (Table 3) vs. individual fairness (Table 4). Within the category of *group fairness* metrics, we furthermore mainly distinguish between the types of *utility* (benefit) in terms of *exposure* and *effectiveness* (Amigó et al. 2023). The metrics listed in Table 4, in contrast, are split into (a) metrics for individual item recommendation scenarios, and (b) metrics for *group recommendation* settings. Exposure and effectiveness can be defined as follows:

- *Exposure* refers to the degree to which an item or group of items is exposed to a user or group of users;
- *Effectiveness* (sometimes called *relevance*) defines the amount to which an item’s exposure is effective, i.e., corresponds to the user’s preferences.

Different stakeholders in recommender systems may be concerned with these two types of utility to varying degrees. For instance, from the perspective of customers, fairness primarily entails an equitable distribution of effectiveness among users, thereby preventing the discrimination of historically disadvantaged groups such as female or black job applicants, for example. In contrast, producers and item providers that seek enhanced visibility are primarily concerned with exposure equity, which should not be punished, for instance, based on producers’ popularity or country.

We note that the popularity of items is a central concept in most metrics that are related to *exposure*. Most commonly, the popularity of an item is assessed in offline experiments by counting the number of observed interactions for each item in the training data. Moreover, various work assume that there is a trade-off between different evaluation objectives: customer fairness, provider fairness, and overall system accuracy. Thus, some metrics in the literature are designed against the background of such potential trade-offs.

Discussion The 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 (Jannach and Jugovac 2019). In fairness research, the corresponding questions are if users would actually consider the recommendations fairer or if a fairness-aware algorithm would lead to the different behavior of the users. Unfortunately, research that involves humans is very rare. An example of a work that considers the effects of fair rankings can be found in Sühr et al. (2021), where mixed effects were observed in the context of job recommendation, accounting for gender biases and the impact of job context, candidate profiles, and employer inherent biases, revealing that fair algorithms are useful unless employers evidence strong gender preferences.

Another potential issue of the metrics used is that they may be a strong oversimplification 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

¹³ We note that in these tables we only provide individual examples of works that used a particular metric.

Table 3 Selected types of evaluation metrics used for *group fairness* scenarios*Metrics used for measuring Exposure*

Popularity of recommended items	Different measures are used in the literature to quantify the popularity of the items in a given list of recommendations, e.g., the Average Recommendation Popularity (ARP) Abdollahpouri et al. (2019a) or the PCOUNT measure in Borges and Stefanidis (2021). The assumption is that recommending less popular items increases fairness, see also Deldjoo et al. (2021b)
Deviation from popularity-ranked list	In Borges and Stefanidis (2021), the authors propose a metric to assess the popularity bias of a list inspired by the Normalized Cumulative Gain (NDCG) metric. The popularity bias is assessed by comparing a given top-n recommendation list with a list that is ranked by popularity. Lists which differ more strongly from a pure popularity-ranked list are considered to be fairer
Proportion of less popular items in recommendations	Different metrics in the literature assess the number of less popular (long-tail) items in the top-n recommendations as a fairness indicator. These metrics are called Average Percentage of Long Tail Items (APLT) in Abdollahpouri et al. (2019a) or Popularity Rate in Ge et al. (2021). Such metrics are commonly based on some pre-defined threshold to distinguish long-tail items from other
Disparate exposure of provider groups	In Boratto et al. (2021b), the authors compare how often the items of a certain group of item providers are recommended relative to the proportion of items of this provider group in the catalog. This measure is used to assess what the authors term “disparate visibility”. A variation of this measure, “disparate exposure”, also includes a positional decay, see also (Gómez et al. 2021). The underlying fairness assumption is that items of a minority group of providers should be recommended to users proportional to their representation
Individual provider exposure	Different exposure-based metrics were proposed which assume that items from the same provider belong to the same group. In Wu et al. (2021b), the variance of the distribution of group-level exposures is used, whereas in Patro et al. (2020) an entropy-like measure is used; in both cases, a lower value evidences less inequality and, hence, more fairness. In Patro et al. (2020) another metric is defined based on a minimum exposure requirement (i.e., each product must be assigned to a minimum number of distinct customers) to measure the fraction of <i>satisfied</i> producers
Variance of provider exposure	Also Wu et al. (2021b) base their fairness assessments on the exposure of the items of providers relative to the number (and quality) of their items in the catalog (as in Boratto et al. (2021b)). The final fairness judgment for a recommender system is however then made by considering the <i>variance</i> of exposures across providers (groups), where lower variance indicates higher fairness
Ranking-based Statistical Parity (RSP)	Zhu et al. (2020b) propose to assess if items of different provider groups have the same probability to be contained in the top- <i>k</i> recommendation lists of users. A system is considered fair if it ensures statistical parity, i.e., when the probability distributions of being ranked (exposed in) in top- <i>k</i> lists are comparable for different groups
Divergence of exposure probabilities	In Dash et al. (2021), Dash et al. aim to assess the probability of exposure for “sponsored” recommendations compared to “organic” recommendations on e-commerce marketplaces. To that purpose they compute the Kullback–Leibler divergence of the distributions, which they estimate based on different factors. A system is considered fair and not exhibiting exposure bias when the divergence is close to zero

Table 3 continued

Concentration on a subset of items	A number of works, e.g., Ge et al. (2021), use the Gini index to assess to what extent a recommender system has a tendency to focus on a limited set of items. Such a concentration on a subset of the items in the catalog may lead to an overproportional, and thus unfair, exposure of some items. The Gini index is a number between 0 and 1, which is traditionally used to quantify inequalities, e.g., in terms of income in a society. A higher Gini index means higher concentration. We however note that this not necessarily means that the concentration is on popular items (which is however usually the case in practice)
<i>Metrics used for measuring Effectiveness^a</i>	
Difference between group's utility	The simplest way to evaluate group fairness is to calculate the <i>difference</i> (typically in an absolute sense) in the <i>average</i> performance of group members where groups are defined based on the protected attributes; here the performance can be quantified using ranking-aware (e.g., NDCG), or rating-based measures (e.g., RMSE). This concept is used to quantify group fairness in a number of publications under several titles, including mean Absolute Difference Zhu et al. (2018b); Deldjoo et al. (2021a, b), or user-oriented group fairness (UGF) Li et al. (2021a), and even Negative bias Misztal-Radecka and Indurkha (2021), where the latter calculates the difference between a performance metric (e.g., NDCG) for a user segment and all other users. It should be highlighted that this metric can be utilized to measure producers' exposure fairness, see e.g., Deldjoo et al. (2021a)
Relevance disparity	This metric was introduced along with "Disparate exposure of provider groups" from above in Boratto et al. (2021b). Essentially, this paper examines the same disparity on the producer-side, but with relevance as the underlying utility. The authors note that a disparity in relevance values might not necessarily imply that the minority group is discriminated based on its exposure or visibility in the recommendations lists, but it may be exacerbated through continuous recommendation loops.
Prediction error across market segment	The average prediction errors of a fair algorithm are supposed to be similar for different market segments. Thus, in Wan et al. (2020) the authors propose to use statistical significance tests and the F-statistic as a fairness evaluation metric to evaluate a global parity of prediction errors across different consumer-product market segments. Lower values in this approach indicate better rating prediction fairness
Ranking-based equal opportunity (REO)	This metric, again introduced by Zhu et al. (2020b), is similar to RSP presented in the previous Table but is primarily concerned with measuring effectiveness fairness. It quantifies the discrepancy between item groups based on the probability that a relevant item is among the top- <i>k</i> suggestions

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 impor-

Table 3 continued*Other Scenarios*

Two-sided metrics	A number of metrics were proposed that integrate two group fairness criteria, namely consumer effectiveness and consumer exposure. (i) <i>Flexible probabilistic metrics</i> : Some works have presented fairness measurement models that are adaptable to specific scenarios, mostly by comparing the distributions provided by a given system against an ideal (fair) distribution, sometimes called <i>target representation</i> , see (Kirnap et al. 2021; Amigó et al. 2023). Generalized Cross Entropy (Deldjoo et al. 2021a, 2019; Rahmani et al. 2022a) is such a metric that compares those two distributions. Similarly, Kirnap et al. (2021) investigate a variety of divergence-based metrics and target representation types (e.g., based on equity, proportionality to the corpus size, etc.); (ii) <i>Joint multi-sided metrics</i> : another group of fairness metrics eliminates the constraint of comparing against a target representation and evaluates fairness on the basis of statistical independence between user and item groups. Examples include Bias Disparity (Tsintzou et al. 2019; Mansoury et al. 2019; Lin et al. 2019) and Mutual Information (Amigó et al. 2023). Another example is Wu et al. (2021b), where the authors study joint multi-sided fairness evaluation by designing metrics to measure the individual fairness of customers, group fairness of providers, and the overall quality of the recommendation results by measuring the <i>quality-weighted exposure</i> for the provider side and comparing the reduction in individuals' recommendation quality for the consumer side (see individual fairness)
Calibration	The assumption behind calibration metrics is that fair recommendations should not deviate from the historical data of the user, this is exactly what User Popularity Deviation (UPD) (Abdollahpouri et al. 2021) measures in terms of the user's interest towards popular items. ΔGAP (Group Average Popularity) by Wundervald (2021) measures the same, but at the (user) group level
Weighted Proportional Fairness	Inspired by rate control algorithms for communication networks, this metric proposed in Liu et al. (2020) is a generalized Nash solution that seeks equilibrium when allocating items (associated with a category) to users. For this, it solves a constrained maximization problem based on the exposure of each group of items

^a Historically, evaluations of fairness in recommender systems mostly associated “exposure” with “providers” and “effectiveness” with “consumers”, as these utilities are of most interest to these stakeholders. However, other works use less explored scenarios, e.g., (Boratto et al. (2021b); Zhu et al. (2020b)), and examine effectiveness from the provider perspective

tant 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 about why the biases lead to unfairness in a certain application context. The normative claims and underlying assumptions about how and when fairness is defined are missing, in parts leading to the impression that the concept of ‘bias mitigation’ instead of ‘fairness’ should have been used. As noted earlier, a similar observation can be made for papers that assume that calibrating recommendations *per se* leads to fairness. This

Table 4 Selected types of evaluation metrics used for *individual fairness* scenarios

Individual recommendation scenario	
Max individual deviation	Addressing the potential trade-off between fairness and other domain-specific requirements/utilities is important. For instance, in particular applications such as mobile apps for video recommendation, regulating fairness, and improving network gains are both crucial goals that may be at odds. Thus, Giannakas et al. (2021) study the problem of network-friendly recommendation (NFR) focusing on owner/producer satisfaction, as measured by the difference in exposure opportunity provided to a single piece of content (item) between the fair recommender being evaluated and a baseline NFR. Individual fairness depends on the metric calculated based on individual content disparity not exceeding a maximum threshold (worst-case scenario), as indicated by the maximum individual deviation. The authors also apply other aggregation metrics, such as total variation distance and Kullback–Leibler Divergence, which eliminate the constraint for individual content and instead concentrate on the disparity on the provider level (group fairness)
The variance of individual losses	In certain research studies, the same trade-off is handled by assuming that the quality of recommendations will decrease when providers' fair exposure is taken into account, and more importantly, that individual fairness can be measured by <i>reduction</i> of individual user recommendation quality. Therefore, it is possible to define individual unfairness as the differences in user losses and to seek for this individual loss value to be dispersed evenly to each consumer, as measured by the difference. Wu et al. (2021b) employ a rank-based measure (NDCG) as the underlying utility for quantifying an individual's recommendation quality, whereas Rastegarpanah et al. (2019) use the mean squared error over a user's known ratings
The variance of user/item deviation cost	Some works connect the notion of utility with the concept of <i>cost</i> . For example, in Koutsopoulos and Halkidi (2018) state that to guarantee a minimum degree of item coverage, e.g., <i>d</i> -coverage, at least <i>d</i> users must be recommended an item. The items in the recommendation list must be re-ranked in order to ensure an optimal ranking under such constraints. Individual fairness is defined by the cost of deviation from a nominal RS that does not account for item coverage and requires the incurred cost of deviation to be as evenly distributed across items or users as possible
Based on Rawlsian fairness	Under Rawlsian principles (Rawls 2001) of <i>justice as fairness</i> and the <i>difference principle</i> (where only inequalities that work to the advantage of the worst-off are permitted), the Max-min opportunity fairness metric (Zhu et al. 2021) accepts inequalities and aims to maximize the minimum utility of individuals or groups so that no subject is underserved by the model; for this, the average true positive rate of the <i>t</i> % worst-off items is computed, which are the <i>t</i> % items with the lowest true positive rates among all cold start items during testing
Group recommendation scenario	
Aggregating effectiveness metrics on a group basis	Some authors aggregate metrics like NDCG or recall according to the users who belong to the same group. For these aggregations, the minimal value in a group or the ratio between minimal and maximal values have been used to quantify the gap between the least and highest utilities of group members in order to achieve social welfare (Malecek and Peska 2021; Kaya et al. 2020)

Table 4 continued

Other uses of effectiveness metrics	In group recommendation settings, where the recommendations for all the users in a group are combined into the same ranking, effectiveness metrics are used as surrogates of fairness to account for how many users are positively impacted by the recommendation. This is done in a way that higher values (more hits, or relevant recommendations for users) mean fairer recommendation lists. Such an approach was chosen in Xiao et al. (2020) with the average reciprocal hit rank, and in Kaya et al. (2020) with the zero-recall metric, which considers how many users received no relevant recommendations. Hence, lower values indicate fairer situations
Satisfaction	Producing recommendations to a group should be fair when multiple iterations are allowed (sequential recommendation). In this context, the authors of (Stratigi et al. 2020) propose several metrics to account for fairness: the overall satisfaction of a user (average of recommendation quality received by a user on each iteration), overall group satisfaction (average of overall user satisfaction across the group), and group disagreement (difference between maximum and minimum satisfaction values in a group)

can probably not be safely stated in general unless the normative claims are made explicit and fit the goals that are achieved by calibration.

When considering recommendation quality metrics for groups, the assumption is either that different groups should have equal recommendation quality (to treat them all alike) or that there is some justified inequality. The latter case may, for example, arise if some groups are assumed to receive better service, e.g., because they have paid for better service or when the inequality is dependent on the corpus size or the available relevant data (Kirnap et al. 2021; Amigó et al. 2023).

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 this line of argumentation is not easy to follow. In fact, some researchers may argue that the correct mitigation strategy would be to fix the data or change the user interface to elicit more data. It would also be debatable which percentage of performance is acceptable to consider such a tradeoff (un)fair, as is the norm in the discussion around statistical parity. Certainly, there may be scenarios where there are particular protected attributes for which it may be desirable not to have largely varying accuracy levels across the groups. In many of the surveyed papers, no realistic use cases are however given.

In terms of the different notions of fairness, traditionally either *group fairness* or *individual fairness* are studied to address consumer effectiveness and producer exposure. However, recent research also addresses situations involving *mixed* individual and group fairness, such as group item exposure fairness and user-individual effectiveness fairness, see for example (Wu et al. 2021b; Rastegarpanah et al. 2019). In such studies, it is often assumed that when provider exposure is addressed, the quality of the recommendations may diminish. The authors thus define individual unfairness as disparities in *user losses* and demand that the decline in recommendation quality be dispersed equitably across all users. As previously stated, the notion of a *trade-off* between the fairness evaluation objectives and overall system accuracy is prevalent

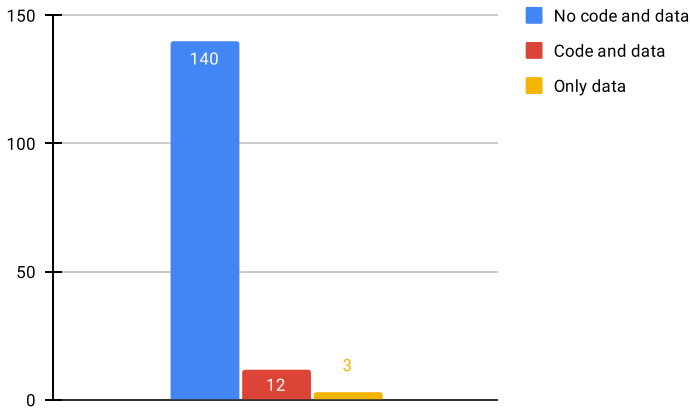


Fig. 11 Level of Reproducibility (Shared Artifacts)

in fairness research, and these demonstrate the need for additional research on multi-sided recommendation fairness.

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 (Masthoff and Delic 2022). 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. It should be noted that Arrow’s impossibility theorem (from *Social Choice Theory*) supports the conclusion that no aggregation strategy will be universally ideal, hence leading again to a potential reason for unfairness in a group. Also in this area researchers, may have been stuck in an abstraction trap (Selbst et al. 2019; Jannach and Bauer 2020) as we have pointed out several oversimplification instances in fairness research, and more (multi-disciplinary) research seems required to understand group recommendation processes, see (Delic et al. 2018) for an observational study in the tourism domain.

Reproducibility The lack of reproducibility can be a major barrier to achieving progress in AI (Gundersen and Kjensmo 2018), and recent studies indicate that limited reproducibility is a substantial issue also in recommender systems research (Cremonesi and Jannach 2021; Bellogín and Said 2021). Figure 11 shows how many of the studied technical papers and artifacts were shared to ensure the reproducibility of the reported experiments.¹⁴ While the level of reproducibility seems to be higher than in general AI (Gundersen and Kjensmo 2018), still for the large majority of the considered works authors did not share any code or data.

¹⁴ The level of reproducibility of research work can be assessed in multiple dimensions, see (Gundersen and Kjensmo 2018). In the context of our work, we limit ourselves to the analysis of certain central artifacts that are publicly shared.

4.6 Landscape overview

Fairness is a multi-faceted subject. In order to provide an encompassing understanding of different fairness dimensions, we have developed a taxonomy that takes different perspectives, as explained in Sect. 3.2, which allows us to describe the landscape of fairness research in recommender systems, as shown in Fig. 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 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) (Kirnap et al. 2021). 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 Kirnap et al. (2021), 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 Sect. 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 Sect. 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. This is connected with the so-called *issue of intersectionality*, which we discuss in some more detail below. 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. Through our survey, we found that there is a balance in the literature between consumer and provider viewpoints. In addition, more recent research in ML has begun to address the issue of *intersectionality* in fairness by building statistical frameworks that account for bias within multiple protected groups, for example, “black women” instead of just “black people” or “women” (Ghosh et al. (2021b); Morina et al. (2019)). An interesting example is presented by Buolamwini and Gebru (2018) where the authors found that commercial facial image classification systems do not show the full distribution of mis-classifications when considering gender and skin type alone, and that darker-skinned women being the most misclassified group, with an accuracy drop of over 30% compared to lighter-skinned men. This aspect has, to the best of our knowledge, been largely overlooked in the recommender fairness research; one exception is the study presented recently by Shen et al. (2023), where such intersectionality between gender (male vs. female) and skin color (black vs. white) fairness was applied to language model-driven conversational recommendation.
- **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 the potential harms of AI-powered systems have received more and more attention in recent years, both in academic research, the tech industry, and within political organizations. Fairness is often considered a central component of what is sometimes called *responsible AI*. These developments can also

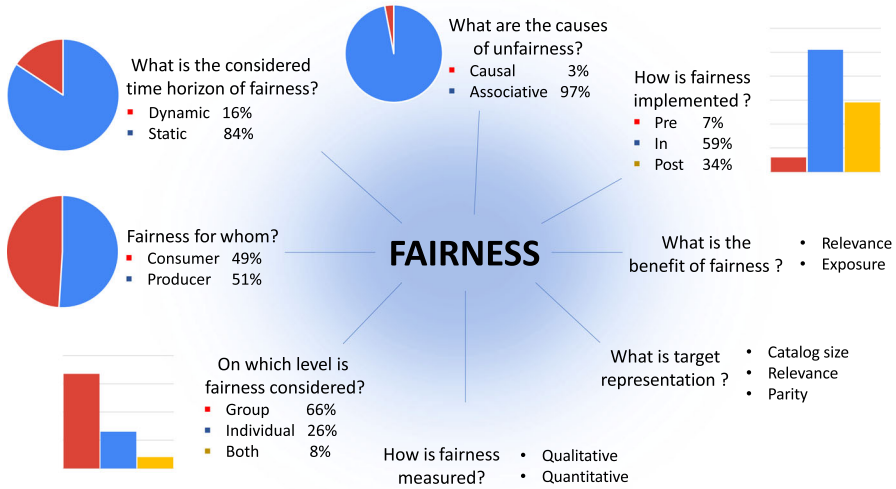


Fig. 12 Taxonomy and landscape

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. Fig. 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. Fig. 2 and Fig. 9). In contrast, only comparably limited research activity seems to take place on topics that go beyond the algorithmic perspective, such as user interfaces and human-in-the-loop approaches, or even beyond computer science (that is applied to AI in general, and recommender systems in particular), such as psychology, economics, or social sciences. 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*” (Narayanan 2018) 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. Figs. 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 its impact on people 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 people are usually the main focus of research, even though these concepts of item fairness are ultimately always related to people (or organizations) in the end, because the item providers are the ones impacted when their items are not recommended.

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. Fig. 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. Fig. 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 missing a clear and detailed discussion of the underlying normative claims that are addressed. 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 an abstraction trap with limited practical impact increases (Selbst et al. 2019; Jannach and Bauer 2020).

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. Less discussion is provided regarding how different level content used in mainstream recommender systems (e.g., user-generated, expert-generated content, and audio) (Moscati et al. 2022; Deldjoo et al. 2021d) are susceptible to the promotion of certain types of biases and unfairness, e.g., audio content could suffer more from an accuracy standpoint but could promote the recommendation of long-term items more effectively. 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 (Ghanem et al. 2022; Rohde et al. 2018; Mladenov et al. 2021; Zhou et al. 2021). 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 realistic and useful definitions for fairness.* As discussed before, there are several definitions for fairness, not only in the RS literature but in ML and AI in general (Olteanu et al. 2019). 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, even from a mathematical point of view (Chouldechova 2017). 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 relevant proposal can be found in Liu et al. (2020), where the authors employ metrics that capture the cumulative reward in a way that combines accuracy and fairness while aiming to improve both. This is a rich area of investigation, open to novel definitions and approaches about how to leverage this tradeoff and whether one dimension should weight more than the other (Friedler et al. 2021; Chouldechova 2017; Kleinberg et al. 2017).

However, this is not the only problem we have identified in our literature review. As stated in Sect. 4.5, the seldom use of user studies and field tests make it very difficult to incorporate user perception (Ferwerda et al. 2023) 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 (Amigó et al. 2023), 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.

- **Challenge 2:** *Building on appropriate data to assess fairness.* As discussed in Sect. 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 (Rastegarpanah et al. 2019). Another, more popular, possibility is to use of simulation instead of real-world datasets. Various recent papers use simulation, sampling techniques (see e.g., the work by Deldjoo et al. (2021b) investigating the impact of data characteristics), and synthetic data to evaluate fairness in search scenarios (Geyik et al. 2019). 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 (Mehrotra et al. 2018).
- **Challenge 3:** *Understanding fairness in reciprocal settings.* Maintaining the utility of stakeholders in reciprocal settings is a new notion of fairness (Xia et al. 2019), 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, see (Koprinska and Yacef 2015) for a survey on people-to-people recommender systems. In the former work, Xia et al. 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 dispar-

ity of service, (b) the similarity of mutual preference, and (c) the equilibrium of demand and supply, that is, by balancing the demand of buyers and the supply of sellers.

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. In fact, going beyond these notions of equilibrium for reciprocal settings, such as cooperative behaviors and non-zero sum games, would require digging further into game theory and related areas, which would be potential avenues for future research.

- **Challenge 4: Fairness auditing.** As stated in Koshiyama et al. (2022), 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 considered in that process. As an example, we want to highlight that the authors from Krafft et al. (2020) 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. Because of this, we argue that, in order to be practical and potentially address this challenge, such requirements should be enforced from higher levels or even policies, otherwise companies may not embrace this type of accountability.

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 (Miller 2019). 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 scholars from fields outside computer science, such as psychology and social science.

Acknowledgements The authors thank the reviewers for their thoughtful comments and suggestions.

Author contributions YD: conceptualization, methodology, investigation, writing, review & editing DJ: conceptualization, methodology, investigation, writing, review & editing AB: conceptualization, methodology, investigation, writing, review & editing AD: investigation DZ: investigation All authors contributed to the article and approved the submitted version.

Funding Open access funding provided by Politecnico di Bari within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abdollahpouri, H., Burke, R.: 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), CEUR Workshop Proceedings, vol. 2440, (2019)
- Abdollahpouri, H., Burke, R., Mobasher, B.: Managing popularity bias in recommender systems with personalized re-ranking. In: Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, pp. 413–418, (2019a)
- Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.: 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, (2019b)
- Abdollahpouri, H., Adomavicius, G., Burke, R., Guy, I., Jannach, D., Kamishima, T., Krasnodebski, J., Pizzato, L.: Multistakeholder recommendation: Survey and research directions. *User Model. User-Adap. Inter.* **30**, 127–158 (2020)
- Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B.: The connection between popularity bias, calibration, and fairness in recommendation. In: Fourteenth ACM Conference on Recommender Systems, pp. 726–731, (2020b)
- Abdollahpouri, H., Mansoury, M., Burke, R., Mobasher, B., Malthouse, E.C.: 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, pp. 119–129, (2021)
- Adomavicius, G., Tuzhilin, A.: Context-aware recommender systems. In: Ricci F, Rokach L, Shapira B (eds) *Recommender Systems Handbook*, pp. 191–226, (2015)
- Adomavicius, G., Jannach, D., Leitner, S., Zhang, J.: Understanding longitudinal dynamics of recommender systems with agent-based modeling and simulation. In: SimuRec Workshop at ACM RecSys 2021, (2021)
- Amigó, E., Deldjoo, Y., Mizzaro, S., Bellogín, A.: A unifying and general account of fairness measurement in recommender systems. *Inf. Process. Manag.* **60**(1), 103115 (2023)
- Anelli, V.W., Belli, L., Deldjoo, Y., Di Noia, T., Ferrara, A., Narducci, F., Pomo, C.: Pursuing privacy in recommender systems: the view of users and researchers from regulations to applications. In: Fifteenth ACM Conference on Recommender Systems, pp. 838–841, (2021)

- Anelli, V.W., Deldjoo, Y., Di Noia, T., Malitesta, D., Paparella, V., Pomo, C.: Auditing consumer- and producer-fairness in graph collaborative filtering. In: Proceedings ECIR '23, (2023)
- Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias. In: Ethics of Data and Analytics, Auerbach Publications, pp. 254–264, (2016)
- Ashokan, A., Haas, C.: Fairness metrics and bias mitigation strategies for rating predictions. *Inf. Process. Manag.* **58**(5), 102646 (2021)
- Baeza-Yates, R.: Bias on the web. *Commun. ACM* **61**(6), 54–61 (2018)
- Barocas, S., Hardt, M., Narayanan, A.: Fairness and Machine Learning. fairmlbook.org, (2019), <http://www.fairmlbook.org>
- Bellogín, A., Said, A.: Improving accountability in recommender systems research through reproducibility. *User Model User Adapt Interact* **31**(5), 941–977 (2021)
- Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E.H., Goodrow, C.: Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, pp. 2212–2220, (2019)
- Binns, R.: On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, pp. 514–524, (2020)
- Bobadilla, J., Lara-Cabrera, R., González-Prieto, Á., Ortega, F.: Deepfair: deep learning for improving fairness in recommender systems. *Int. J. Interact. Multim. Artif. Intell.* **6**(6), 86–94 (2021)
- Boratto, L., Fenu, G., Marras, M.: Connecting user and item perspectives in popularity debiasing for collaborative recommendation. *Inf. Process. Manag.* **58**(1), 102387 (2021)
- Boratto, L., Fenu, G., Marras, M.: Interplay between upsampling and regularization for provider fairness in recommender systems. *User Model User Adapt Interact* **31**(3), 421–455 (2021)
- Borges, R., Stefanidis, K.: On mitigating popularity bias in recommendations via variational autoencoders. In: SAC '21: The 36th ACM/SIGAPP Symposium on Applied Computing, pp. 1383–1389, (2021)
- Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on Fairness, Accountability and Transparency, PMLR (2018)
- Burke, R.: Multisided fairness for recommendation. In: 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017), (2017)
- Burke, R., Sonboli, N., Mansoury, M., Ordoñez-Gauger, A.: Balanced neighborhoods for fairness-aware collaborative recommendation, (2017)
- Burke, R., Sonboli, N., Ordoñez-Gauger, A.: Balanced neighborhoods for multi-sided fairness in recommendation. In: Conference on Fairness, Accountability and Transparency, FAT 2018, Proceedings of Machine Learning Research, vol. 81, pp. 202–214, (2018)
- Chakraborty, A., Messias, J., Benevenuto, F., Ghosh, S., Ganguly, N., Gummadi, K.P.: Who Makes Trends? Understanding Demographic Biases in Crowdsourced Recommendations. In: Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, pp. 22–31 (2017)
- Chakraborty, A., Patro, G.K., Ganguly, N., Gummadi, K.P., Loiseau, P.: Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, pp. 129–138, (2019)
- Chaudhari, H.A., Lin, S., Linda, O.: A general framework for fairness in multistakeholder recommendations. (2020), [arXiv:2009.02423](https://arxiv.org/abs/2009.02423)
- Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., He, X.: Bias and debias in recommender system: a survey and future directions. *ACM Trans. Inf. Syst.* (2022)
- Chouldechova, A.: Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* **5**(2), 153–163 (2017)
- Cooper, A.F.: Where is the normative proof? assumptions and contradictions in ML fairness research. (2020), CoRR [arXiv:2010.10407](https://arxiv.org/abs/2010.10407)
- Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: a critical review of fair machine learning. (2018), CoRR [arXiv:1808.00023](https://arxiv.org/abs/1808.00023)
- Cornacchia, G., Narducci, F., Ragone, A.: A general model for fair and explainable recommendation in the loan domain. 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), (2021)
- Council, N.R., et al.: Measuring Racial Discrimination. National Academies Press, London (2004)

- Cremonesi, P., Jannach, D.: Progress in recommender systems research: crisis? What crisis? *AI Mag.* **42**(3), 43–54 (2021)
- da Silva, D.C., Manzato, M.G., Durão, F.A.: Exploiting personalized calibration and metrics for fairness recommendation. *Expert Syst. Appl.* **181**, 115112 (2021)
- Dash, A., Chakraborty, A., Ghosh, S., Mukherjee, A., Gummadi, K.P.: 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*, pp. 873–884, (2021)
- Deldjoo, Y., Anelli, V.W., Zamani, H., Kouki, A.B., Noia, T.D.: Recommender systems fairness evaluation via generalized cross entropy. In: *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, (2019)
- Deldjoo, Y., Di Noia, T., Merra, F.A.: Adversarial machine learning in recommender systems (aml-recsys). In: *Proceedings of the 13th International Conference on Web Search and Data Mining*, pp. 869–872, (2020)
- Deldjoo, Y., Anelli, V.W., Zamani, H., Bellogin, A., Di Noia, T.: A flexible framework for evaluating user and item fairness in recommender systems. *User Modeling and User-Adapted Interaction* pp. 1–47, (2021a)
- Deldjoo, Y., Bellogin, A., Di Noia, T.: Explaining recommender systems fairness and accuracy through the lens of data characteristics. *Inf. Process. Manag.* **58**(5), 102662 (2021)
- Deldjoo, Y., Noia, T.D., Merra, F.A.: A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. *ACM Comput. Surv.* **54**(2), 1–38 (2021)
- Deldjoo, Y., Schedl, M., Knees, P.: Content-driven music recommendation: evolution, state of the art, and challenges. (2021d), arXiv preprint [arXiv:2107.11803](https://arxiv.org/abs/2107.11803)
- Deldjoo, Y., Nazary, F., Ramisa, A., McAuley, J., Pellegrini, G., Bellogin, A., Di Noia, T.: A review of modern fashion recommender systems. *ACM Comput. Surv.* (2023)
- Delic, A., Neidhardt, J., Nguyen, T.N., Ricci, F.: An observational user study for group recommender systems in the tourism domain. *J. Inf. Technol. Tour.* **19**(1–4), 87–116 (2018)
- Di Noia, T., Tintarev, N., Fatourou, P., Schedl, M.: Recommender systems under European AI regulations. *Commun. ACM* **65**(4), 69–73 (2022)
- Dong, Q., Xie, S., Li, W.: User-item matching for recommendation fairness. *IEEE Access* **9**, 130389–130398 (2021)
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.S.: Fairness through awareness. In: Goldwasser S (ed) *Innovations in Theoretical Computer Science 2012*, pp. 214–226, (2012)
- Edizel, B., Bonchi, F., Hajian, S., Panisson, A., Tassa, T.: Fairecsys: mitigating algorithmic bias in recommender systems. *Int. J. Data Sci. Anal.* **9**(2), 197–213 (2020)
- Ekstrand, M.D., Das, A., Burke, R., Diaz, F.: Fairness in information access systems. *Found. Trends Inf. Retr.* **16**(1–2), 1–177 (2022)
- Elahi, M., Jannach, D., Skjærven, L., Knudsen, E., Sjøvaag, H., Tolonen, K., Holmstad, Ø., Pipkin, I., Thronsen, E., Stenbom, A., Fiskerud, E., Oesch, A., Vredenberg, L., Trattner, C.: Towards responsible media recommendation. *AI Ethics* **2**, 103–114 (2022)
- Farnadi, G., Kouki, P., Thompson, S.K., Srinivasan, S., Getoor, L.: A fairness-aware hybrid recommender system, (2018). [arXiv:1809.09030](https://arxiv.org/abs/1809.09030)
- Felfernig, A., Boratto, L., Stettinger, M., Tkali, M.: *Group Recommender Systems: An Introduction*. Springer, Berlin (2018)
- Ferraro, A.: Music cold-start and long-tail recommendation: bias in deep representations. In: Bogers T, Said A, Brusilovsky P, Tikk D (eds) *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019*, pp. 586–590, (2019)
- Ferwerda, B., Ingesson, E., Berndl, M., Schedl, M.: I Don't Care How Popular You Are! Investigating Popularity Bias From a User's Perspective. In: *Proceedings of the 8th ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 2023)*, ACM, Austin, USA, (2023)
- Friedler, S.A., Scheidegger, C., Venkatasubramanian, S.: The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM* **64**(4), 136–143 (2021)
- Friedman, A., Knijnenburg, B.P., Vanhecke, K., Martens, L., Berkovsky, S.: Privacy aspects of recommender systems. In: *Recommender Systems Handbook*, Springer, pp. 649–688, (2015)
- Friedman, B., Nissenbaum, H.: Bias in computer systems. *ACM Trans. Inf. Syst.* **14**(3), 330–347 (1996)
- Fu, Z., Xian, Y., Gao, R., Zhao, J., Huang, Q., Ge, Y., Xu, S., Geng, S., Shah, C., Zhang, Y., de Melo, G.: 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, pp. 69–78, (2020)
- Ge, Y., Liu, S., Gao, R., Xian, Y., Li, Y., Zhao, X., Pei, C., Sun, F., Ge, J., Ou, W., Zhang, Y.: Towards long-term fairness in recommendation. In: WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, pp. 445–453, (2021)
- Geyik, S.C., Ambler, S., Kenthapadi, K., Karypis, G.: 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, pp. 2221–2231, (2019)
- Ghanem, N., Leitner, S., Jannach, D.: Balancing consumer and business value of recommender systems: a simulation-based analysis. *E-Commerce Research and Applications* forthcoming, (2022)
- Gharahighihi, A., Vens, C., Pliakos, K.: Fair multi-stakeholder news recommender system with hypergraph ranking. *Inf. Process. Manag.* **58**(5), 102663 (2021)
- Ghosh, A., Dutt, R., Wilson, C.: When fair ranking meets uncertain inference. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1033–1043, (2021a)
- Ghosh, A., Genuit, L., Reagan, M.: Characterizing intersectional group fairness with worst-case comparisons. In: *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, PMLR, pp. 22–34, (2021b)
- Giannakas, T., Sermpezis, P., Giovanidis, A., Spyropoulos, T., Arvanitakis, G.: Fairness in network-friendly recommendations. In: 22nd IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, WoWMoM 2021, pp. 71–80, (2021)
- Gómez, E., Zhang, C.S., Boratto, L., Salamó, M., Marras, M.: The winner takes it all: Geographic imbalance and provider (un)fairness in educational recommender systems. In: SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1808–1812, (2021)
- Gorantla, S., Deshpande, A., Louis, A.: On the problem of underranking in group-fair ranking. In: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, Proceedings of Machine Learning Research, vol. 139, pp. 3777–3787, (2021)
- Grgic-Hlaca, N., Zafar, M.B., Gummadi, K.P., Weller, A.: The case for process fairness in learning: Feature selection for fair decision making. In: *NIPS Symposium on Machine Learning and the Law*, Barcelona, Spain, vol. 1, pp. 2, (2016)
- Gunawardana, A., Shani, G., Yogev, S.: Evaluating recommender systems. In: Rokach, L., Shapira, B., Ricci, F. (eds.) *Recommender Systems Handbook*, pp. 547–601. Springer, Berlin (2022)
- Gundersen, O.E., Kjensmo, S.: 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), pp. 1644–1651, (2018)
- Gupta, A., Johnson, E., Payan, J., Roy, A.K., Kobren, A., Panda, S., Tristan, J.B., Wick, M.: 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, pp. 454–462, (2021)
- Hao, Q., Xu, Q., Yang, Z., Huang, Q.: Pareto optimality for fairness-constrained collaborative filtering. In: *MM '21: ACM Multimedia Conference*, ACM, pp. 5619–5627, (2021)
- Harper, F.M., Konstan, J.A.: The MovieLens datasets: history and context. *ACM Trans. Interact. Intell. Syst.* **5**(4), (2015)
- Htun, N.N., Lecluse, E., Verbert, K.: Perception of fairness in group music recommender systems. In: 26th International Conference on Intelligent User Interfaces, pp. 302–306, (2021)
- Jannach, D., Adomavicius, G.: Price and profit awareness in recommender systems. In: Proceedings of the ACM RecSys 2017 Workshop on Value-Aware and Multi-Stakeholder Recommendation, (2017)
- Jannach, D., Bauer, C.: Escaping the McNamara fallacy: towards more impactful recommender systems research. *AI Mag.* **41**(4), 79–95 (2020)
- Jannach, D., Jugovac, M.: Measuring the business value of recommender systems. *ACM TMIS* **10**(4), 1–23 (2019)
- Jannach, D., Zanker, M., Felfernig, A., Friedrich, G.: *Recommender Systems-An Introduction*. Cambridge University Press, Cambridge (2010)
- Jannach, D., Zanker, M., Ge, M., Gröning, M.: 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), pp. 76–87, (2012)

- Jannach, D., Lерche, L., Kamehkhosh, I., Jugovac, M.: What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Model. User-Adap. Inter.* **25**(5), 427–491 (2015)
- Jannach, D., Resnick, P., Tuzhilin, A., Zanker, M.: Recommender systems-beyond matrix completion. *Commun. ACM* **59**(11), 94–102 (2016)
- Jannach, D., Pu, P., Ricci, F., Zanker, M.: Recommender systems: past, present, future. *AI Mag.* **42**(3), 3–6 (2021)
- Jugovac, M., Jannach, D., Lерche, L.: Efficient optimization of multiple recommendation quality factors according to individual user tendencies. *Expert Syst. Appl.* **81**, 321–331 (2017)
- Kaya, M., Bridge, D., Tintarev, N.: Ensuring Fairness in Group Recommendations by Rank-Sensitive Balancing of Relevance, pp. 101–110, (2020)
- Kirnap, Ö., Diaz, F., Biega, A., Ekstrand, M.D., Carterette, B., Yilmaz, E.: Estimation of fair ranking metrics with incomplete judgments. In: WWW '21: The Web Conference 2021, pp. 1065–1075, (2021)
- Kitchenham, B.A., Brereton, P., Budgen, D., Turner, M., Bailey, J., Linkman, S.G.: Systematic literature reviews in software engineering—a systematic literature review. *Inf. Softw. Technol.* **51**(1), 7–15 (2009)
- Kleinberg, J.M., Mullainathan, S., Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. In: 8th Innovations in Theoretical Computer Science Conference, ITCS, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, LIPIcs, vol. 67, pp. 43:1–43:23, (2017)
- Koprinska, I., Yacef, K.: People-to-people reciprocal recommenders. In: *Recommender Systems Handbook*, Springer, pp. 545–567, (2015)
- Koshiyama, A.S., Kazim, E., Treleaven, P.C.: Algorithm auditing: Managing the legal, ethical, and technological risks of artificial intelligence, machine learning, and associated algorithms. *Computer* **55**(4), 40–50 (2022)
- Koutsopoulos, I., Halkidi, M.: 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, pp. 912–918, (2018)
- Krafft, T.D., Hauer, M.P., Zweig, K.A.: 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, pp. 27–34, (2020)
- Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. In: *Advances in Neural Information Processing Systems*, vol. 30, (2017)
- Li, Y., Chen, H., Fu, Z., Ge, Y., Zhang, Y.: User-oriented fairness in recommendation. In: *Proceedings of The Web Conference 2021, WWW '21*, pp. 624–632, (2021a)
- Li, Y., Chen, H., Xu, S., Ge, Y., Zhang, Y.: Towards personalized fairness based on causal notion. In: 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, pp. 1054–1063, (2021b)
- Li, Y., Ge, Y., Zhang, Y.: Tutorial on fairness of machine learning in recommender systems. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2654–2657, (2021c)
- Li, Y., Chen, H., Xu, S., Ge, Y., Tan, J., Liu, S., Zhang, Y.: Fairness in recommendation: a survey. (2022), [CoRR arXiv:2205.13619](https://arxiv.org/abs/2205.13619)
- Lin, C., Liu, X., Xv, G., Li, H.: Mitigating sentiment bias for recommender systems. In: *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 31–40, (2021)
- Lin, K., Sonboli, N., Mobasher, B., Burke, R.: 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 Recommender Systems (RecSys 2019), CEUR Workshop Proceedings*, vol. 2440, (2019)
- Lin, K., Sonboli, N., Mobasher, B., Burke, R.: Calibration in collaborative filtering recommender systems: a user-centered analysis. In: *HT '20: 31st ACM Conference on Hypertext and Social Media*, pp. 197–206, (2020)
- Liu, W., Liu, F., Tang, R., Liao, B., Chen, G., Heng, P.: 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, pp. 155–167, (2020)

- Malecek, L., Peska, L.: Fairness-preserving group recommendations with user weighting. In: Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization, pp. 4–9, (2021)
- Mansoury, M., Mobasher, B., Burke, R., Pechenizkiy, M.: 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)
- Masthoff, J., Delic, A.: Group recommender systems: beyond preference aggregation. In: Rokach, L., Shapira, B., Kantor, P., Ricci, F. (eds.) Recommender Systems Handbook. Springer, Berlin (2022)
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. *ACM Comput Surv* 54(6), (2021)
- Mehrotra, R., McInerney, J., Bouchard, H., Lalmas, M., Diaz, F.: 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, pp 2243–2251, (2018)
- Melchiorre, A.B., Rekabsaz, N., Parada-Cabaleiro, E., Brandl, S., Lesota, O., Schedl, M.: Investigating gender fairness of recommendation algorithms in the music domain. *Inf. Process. Manag.* 58(5), 102666 (2021)
- Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* 267, 1–38 (2019)
- Misztal-Radecka, J., Indurkha, B.: Bias-aware hierarchical clustering for detecting the discriminated groups of users in recommendation systems. *Information Processing & Management* 58(3), 102519 (2021)
- Mladenov, M., Hsu, C., Jain, V., Ie, E., Colby, C., Mayoraz, N., Pham, H., Tran, D., Vendrov, I., Boutilier, C.: RecSim NG: toward principled uncertainty modeling for recommender ecosystems. (2021), CoRR [arXiv:2103.08057](https://arxiv.org/abs/2103.08057)
- Morina, G., Oliinyk, V., Waton, J., Marusic, I., Georgatzis, K.: Auditing and achieving intersectional fairness in classification problems. (2019), arXiv preprint [arXiv:1911.01468](https://arxiv.org/abs/1911.01468)
- Moscati, M., Parada-Cabaleiro, E., Deldjoo, Y., Zangerle, E., Schedl, M.: Music4all-onion. a large-scale multi-faceted content-centric music recommendation dataset. In: Proceedings of the 31th ACM International Conference on Information & Knowledge Management (CIKM'22), (2022)
- Mulligan, D.K., Kroll, J.A., Kohli, N., Wong, R.Y.: This thing called fairness: Disciplinary confusion realizing a value in technology. *Proc ACM Hum Comput Interact* 3(CSCW), 119:1-119:36 (2019)
- Naghiaei, M., Rahmani, H.A., Deldjoo, Y.: CPFair: Personalized Consumer and Producer Fairness Re-ranking for Recommender Systems. In: SIGIR '22SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, (2022)
- Narayanan, A.: 21 definitions of fairness and their politics. Tutorial at FAT* 2018, (2018)
- Ntoutsis, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejd, W., Vidal, M., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernández, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T., Staab, S.: Bias in data-driven artificial intelligence systems - an introductory survey. *WIREs Data Mining Knowl Discov* 10(3), (2020)
- Nunes, I., Jannach, D.: A systematic review and taxonomy of explanations in decision support and recommender systems. *User-Modeling and User-Adapted Interaction* 27(3–5), 393–444 (2017)
- Oh, J., Park, S., Yu, H., Song, M., Park, S.T.: Novel recommendation based on personal popularity tendency. In: ICDM '11, pp 507–516, (2011)
- Olteanu, A., Castillo, C., Diaz, F., Kiciman, E.: Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers Big Data* 2, 13 (2019)
- Patro, G.K., Biswas, A., Ganguly, N., Gummadi, K.P., Chakraborty, A.: Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In: WWW '20: The Web Conference 2020, pp 1194–1204, (2020)
- Pedreshi, D., Ruggieri, S., Turini, F.: Discrimination-aware data mining. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, p 560–568, (2008)
- Pessach, D., Shmueli, E.: A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55(3), 1–44 (2022)

- Pitoura, E., Stefanidis, K., Koutrika, G.: Fairness in rankings and recommendations: an overview. *VLDB J.* **31**(3), 431–458 (2022)
- Qiu, R., Wang, S., Chen, Z., Yin, H., Huang, Z.: CausalRec: Causal Inference for Visual Debiasing in Visually-Aware Recommendation. In: *MM '21: ACM Multimedia Conference*, pp 3844–3852, (2021)
- Rahmani, H.A., Deldjoo, Y., di Noia, T.: The role of context fusion on accuracy, beyond-accuracy, and fairness of point-of-interest recommendation systems. *Expert Systems with Applications* p 117700, (2022a)
- Rahmani, H.A., Deldjoo, Y., Tourani, A., Naghiaei, M.: The unfairness of active users and popularity bias in point-of-interest recommendation. In: *Advances in Bias and Fairness in Information Retrieval - Third International Workshop, BIAS 2022*, Springer, Communications in Computer and Information Science, vol 1610, pp 56–68, (2022b)
- Rahmani, H.A., Naghiaei, M., Dehghan, M., Aliannejadi, M.: Experiments on generalizability of user-oriented fairness in recommender systems. In: *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp 2755–2764, (2022c)
- Rahmani, H.A., Naghiaei, M., Tourani, A., Deldjoo, Y.: Exploring the impact of temporal bias in point-of-interest recommendation. In: *Proceedings of the 16th ACM Conference on Recommender Systems*, (2022d)
- Rastegarpanah, B., Gummadi, K.P., Crovella, M.: 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*, pp 231–239, (2019)
- Rawls, J.: *Justice as fairness: A restatement*. Harvard University Press (2001)
- Riederer, C., Chaintreau, A.: The price of fairness in location based advertising. In: *FATREC'17*, (2017)
- Rohde, D., Bonner, S., Dunlop, T., Vasile, F., Karatzoglou, A.: Recogym: A reinforcement learning environment for the problem of product recommendation in online advertising. *arXiv preprint arXiv:1808.00720*, (2018)
- Schelenz, L.: 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*, pp 404–410, (2021)
- Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and abstraction in sociotechnical systems. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, p 59–68, (2019)
- Serbos, D., Qi, S., Mamoulis, N., Pitoura, E., Tsaparas, P.: Fairness in package-to-group recommendations. In: *Proceedings of the 26th International Conference on World Wide Web, WWW 2017*, pp 371–379, (2017)
- Seymen, S., Abdollahpouri, H., Malthouse, E.C.: 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)*, CEUR Workshop Proceedings, vol 2959, (2021)
- Shakespeare, D., Porcaro, L., Gómez, E., Castillo, C.: 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)*, CEUR Workshop Proceedings, vol 2697, (2020)
- Shen, T., Li, J., Bouadjenek, M.R., Mai, Z., Sanner, S.: Towards understanding and mitigating unintended biases in language model-driven conversational recommendation. *Information Processing and Management In press*, (2023)
- Shrestha, Y.R., Yang, Y.: Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. *Algorithms* **12**(9), 199 (2019)
- Slokom, M., Hanjalic, A., Larson, M.: Towards user-oriented privacy for recommender system data: A personalization-based approach to gender obfuscation for user profiles. *Information Processing & Management* **58**(6), 102722 (2021)
- Sonboli, N., Burke, R., Mattei, N., Eskandarian, F., Gao, T.: “and the winner is...”: Dynamic lotteries for multi-group fairness-aware recommendation. In: *FACCTRec Workshop: Responsible Recommendation (RecSys '20)*, (2020)
- Sonboli, N., Smith, J.J., Cabral Berenfus, F., Burke, R., Fiesler, C.: Fairness and transparency in recommendation: The users' perspective. In: *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, p 274–279, (2021)

- Srivastava, M., Heidari, H., Krause, A.: Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, pp 2459–2468, (2019)
- Steck, H.: Calibrated recommendations. In: Proceedings of the 12th ACM Conference on Recommender Systems, pp 154–162, (2018)
- Stratigi, M., Kondylakis, H., Stefanidis, K.: Fairness in group recommendations in the health domain. In: 33rd IEEE International Conference on Data Engineering, ICDE 2017, pp 1481–1488, (2017)
- Stratigi, M., Nummenmaa, J., Pitoura, E., Stefanidis, K.: Fair sequential group recommendations. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing, pp 1443–1452, (2020)
- Sühr, T., Hilgard, S., Lakkaraju, H.: Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring, p 989–999, (2021)
- Sun, W., Khenissi, S., Nasraoui, O., Shafto, P.: Debiasing the human-recommender system feedback loop in collaborative filtering. In: Companion of The 2019 World Wide Web Conference, WWW 2019, ACM, pp 645–651, (2019)
- Tintarev, N., Masthoff, J.: Beyond explaining single item recommendations. In: Recommender Systems Handbook, Springer, pp 711–756, (2022)
- Trattner, C., Jannach, D., Motta, E., Meijer, I.C., Diakopoulos, N., Elahi, M., Opdahl, A.L., Tessem, B., Borch, N., Fjeld, M., Øvrelid, L., Smedt, K.D., Moe, H.: Responsible Media Technology and AI: Challenges and Research Directions. *AI and Ethics* **2**, 585–594 (2022)
- Tsintzou, V., Pitoura, E., Tsaparas, P.: 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)
- Verma, S., Rubin, J.: Fairness definitions explained. In: Brun Y, Johnson B, Meliou A (eds) Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, pp 1–7, (2018)
- Verma, S., Gao, R., Shah, C.: Facets of fairness in search and recommendation. In: Bias and Social Aspects in Search and Recommendation - First International Workshop, BIAS 2020, Communications in Computer and Information Science, vol 1245, pp 1–11, (2020)
- Wan, M., Ni, J., Misra, R., McAuley, J.: Addressing marketing bias in product recommendations. In: Proceedings of the 13th International Conference on Web Search and Data Mining, pp 618–626, (2020)
- Wang, C., Wang, K., Bian, A., Islam, R., Keya, K.N., Foulds, J.R., Pan, S.: Do humans prefer debiased AI algorithms? A case study in career recommendation. In: IUI 2022: 27th International Conference on Intelligent User Interfaces, pp 134–147, (2022a)
- Wang, X., Thain, N., Sinha, A., Prost, F., Chi, E.H., Chen, J., Beutel, A.: Practical compositional fairness: Understanding fairness in multi-component recommender systems. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining, pp 436–444, (2021)
- Wang, Y., Ma, W., Zhang, M., Liu, Y., Ma, S.: A survey on the fairness of recommender systems. *ACM TOIS* forthcoming, (2022b)
- Weydemann, L., Sacharidis, D., Werthner, H.: 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, pp 6:1–6:8, (2019)
- Wu, C., Wu, F., Wang, X., Huang, Y., Xie, X.: 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, pp 4462–4469, (2021a)
- Wu, Y., Cao, J., Xu, G., Tan, Y.: 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, pp 1013–1022, (2021b)
- Wundervald, B.D.: Cluster-based quotas for fairness improvements in music recommendation systems. *Int J Multim Inf Retr* **10**(1), 25–32 (2021)
- Xia, B., Yin, J., Xu, J., Li, Y.: We-rec: A fairness-aware reciprocal recommendation based on walrasian equilibrium. *Knowl Based Syst* **182**, (2019)
- Xiao, B., Benbasat, I.: E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Q.* **31**(1), 137–209 (2007)
- Xiao, Y., Pei, Q., Yao, L., Yu, S., Bai, L., Wang, X.: An enhanced probabilistic fairness-aware group recommendation by incorporating social activeness. *J. Netw. Comput. Appl.* **156**, 102579 (2020)

- Yadav, H., Du, Z., Joachims, T.: 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, pp 1044–1053, (2021)
- Yao, S., Huang, B.: Beyond parity: Fairness objectives for collaborative filtering. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17, p 2925–2934, (2017)
- Zafar, M.B., Valera, I., Gomez Rodriguez, M., Gummadi, K.P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In: Proceedings of the 26th International Conference on World Wide Web, pp 1171–1180, (2017)
- Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in ranking, part i: Score-based ranking. *ACM Comput Surv Just Accepted*, (2022a)
- Zehlike, M., Yang, K., Stoyanovich, J.: Fairness in ranking, part ii: Learning-to-rank and recommender systems. *ACM Comput Surv* forthcoming, (2022b)
- Zhang, J., Bareinboim, E.: Fairness in decision-making—the causal explanation formula. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 32, (2018)
- Zhao, Z., Chen, J., Zhou, S., He, X., Cao, X., Zhang, F., Wu, W.: Popularity bias is not always evil: Disentangling benign and harmful bias for recommendation. *IEEE Transactions on Knowledge & Data Engineering* (01):1–13, (2022)
- Zheng, Y., Dave, T., Mishra, N., Kumar, H.: Fairness in reciprocal recommendations: A speed-dating study. In: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, UMAP 2018, pp 29–34, (2018)
- Zhou, M., Zhang, J., Adomavicius, G.: Longitudinal impact of preference biases on recommender systems' performance. *Kelley School of Business* (2021-10), (2021)
- Zhu, Q., Zhou, A., Sun, Q., Wang, S., Yang, F.: 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, pp 171–178, (2018a)
- Zhu, Q., Sun, Q., Li, Z., Wang, S.: FARM: A fairness-aware recommendation method for high visibility and low visibility mobile apps. *IEEE Access* **8**, 122747–122756 (2020)
- Zhu, Z., Hu, X., Caverlee, J.: Fairness-aware tensor-based recommendation. In: Proceedings of the 27th International Conference on Information and Knowledge Management, CIKM 2018, pp 1153–1162, (2018b)
- Zhu, Z., Wang, J., Zhang, Y., Caverlee, J.: Fairness-aware recommendation of information curators. (2018c), [arXiv:1809.03040](https://arxiv.org/abs/1809.03040)
- Zhu, Z., Wang, J., Caverlee, J.: 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, p 449–458, (2020b)
- Zhu, Z., Kim, J., Nguyen, T., Fenton, A., Caverlee, J.: 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, pp 767–776, (2021)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Yashar Deldjoo is a tenure-track assistant professor at the Polytechnic University of Bari in Italy, where his area of specialization is in “multimodal information-seeking systems” and “trustworthy AI.” He earned his Ph.D. from the Polytechnic University of Milan in 2018 with distinction. Dr. Deldjoo's research focuses on incorporating trustworthy AI principles, such as fairness, robustness, privacy, and explainability, into recommendation and machine learning algorithms. Additionally, he has been working on the application of causal inference in ML systems. He is a leading author of a chapter on the recommender system handbook and several surveys in the widely read ACM computing survey journal. He serves as a senior PC and program chair at prominent IR conferences such as SIGIR, The WebConference, CIKM, and EWAF.

Dietmar Jannach is a Professor of Information Systems at the University of Klagenfurt, Austria. He has worked on different areas of artificial intelligence, including recommender systems, model-based diagnosis, and knowledge-based systems. He is the leading author of a textbook on recommender systems and

has authored more than hundred research papers, focusing on the application of artificial intelligence technology to practical problems.

Alejandro Bellogín is Associate Professor at Universidad Autonoma de Madrid (UAM), Spain. Previously, he worked with Prof. Arjen de Vries associated to the Centrum Wiskunde & Informatica under apost-doctoral Marie Curie research grant. In 2012, he completed his PhD under the supervision of Prof. Pablo Castells and Dr. Iván Cantador at UAM. Dr. Bellogín has worked in several areas of user modeling and personalization, including recommender systems, evaluation, and reproducibility, where he has published over 80 conference and journal papers, while being involved as program or organization committees in venues related to these areas, such as RecSys, WWW, and UMAP, among others.

Alessandro Difonzo is an MSc student of Computer Engineering at the Polytechnic of Bari, Italy, with a keen research interest in fairness in recommender systems and simulation and data synthesis. Currently, he is completing his MSc thesis under the guidance of Dr. Yashar Deldjoo, delving into the impact of sampling and data characteristics on recommender systems. Alessandro is also a co-contributor of a well-received UMUAI journal survey on the fairness of recommender systems.

Dario Zabzonelli is an MSc student of Computer Engineering at the Polytechnic of Bari, Italy, and has contributed to a survey on the topic of fairness in recommender systems, published in the UMUAI journal.