Artificial Intelligence Review (2024) 57:101
https://doi.org/10.1007/s10462-023-10663-5

Consumer-side fairness in recommender systems: a systematic survey of methods and evaluation

Bjørnar Vassøy[1] · Helge Langseth[1]

Accepted: 19 December 2023 / Published online: 29 March 2024

Abstract

In the current landscape of ever-increasing levels of digitalization, we are facing major challenges pertaining to data volume. Recommender systems have become irreplaceable both for helping users navigate the increasing amounts of data and, conversely, aiding providers in marketing products to interested users. Data-driven models are susceptible to data bias, materializing in the bias influencing the models' decision-making. For recommender systems, such issues are well exemplified by occupation recommendation, where biases in historical data may lead to recommender systems relating one gender to lower wages or to the propagation of stereotypes. In particular, consumer-side fairness, which focuses on mitigating discrimination experienced by users of recommender systems, has seen a vast number of diverse approaches. The approaches are further diversified through differing ideas on what constitutes fair and, conversely, discriminatory recommendations. This survey serves as a systematic overview and discussion of the current research on consumer-side fairness in recommender systems. To that end, a novel taxonomy based on high-level fairness definitions is proposed and used to categorize the research and the proposed fairness evaluation metrics. Finally, we highlight some suggestions for the future direction of the field.

Keywords Recommender systems · Fairness · Survey · Consumer-side fairness

1 Introduction

Recommender systems have become integral parts of modern digital society. An exponential increase in data poses significant challenges to users and consumers, who cannot feasibly sift through everything to find what they are looking for. Recommender systems help mitigate these challenges by capturing their users' preferences and presenting them with prioritized options. Thus, recommender systems have seen widespread application in e-commerce, multimedia platforms, and social networks. Their tactical relevance in

✉ Bjørnar Vassøy
bjornar.vassoy@ntnu.no

Helge Langseth
helge.langseth@ntnu.no

[1] Department of Computer Science, Norwegian University of Science and Technology, Høgskoleringen 1, 7034 Trondheim, Trøndelag, Norway

the industry has led to a high degree of cooperation between the industry and academia in further developing the field (Jannach et al. 2010; Ricci et al. 2022a).

In recent years, the notion of fairness in machine learning has steadily gained attention. High-profile cases have succeeded in bringing the topic to the general public's attention, like the analysis performed by ProPublica suggesting the presence of racial bias in the COMPAS system used for predicting the likelihood of recidivism of inmates (ProPublica 2016). Subsequently, fairness challenges have also been identified for recommender systems, and the works of Burke et al. (2018) formalized the presence of multi-stakeholder fairness dynamics mirroring the multi-stakeholder nature of recommender systems. Provider stakeholders may take issue if their products are disproportionally less exposed than similar popular products. Career seekers may feel discriminated against if they are predominantly recommended careers that are stereotypically and historically associated with their gender. An increased focus on fairness in recommender systems is not only ethically beneficial for society as a whole but also helps the actors applying them in satisfying an increasingly fairness-aware user base and retaining good relations and cooperation with providers (Ekstrand et al. 2022).

While provider-side fairness research has a dominant subgroup in research pursuing popularity bias, which is the notion of disproportional amounts of attention given to popular items, consumer-side fairness research has a greater focus on group-based fairness relating to demographic information of the users, i.e., making sure that users are not discriminated against based on aspects such as race, gender, or age. Despite the focus on a specific high-level fairness setting, consumer-side fairness in recommender systems displays a high degree of variation in approaches. The approaches for introducing fairness awareness take place in all parts of the recommender system pipeline, span most established and upcoming model architectures, and are designed to satisfy various fairness definitions. Some models opt for adjusting recommendations post hoc, others modify the input data directly, while others still explicitly model the fairness-awareness. Fairness has been incorporated through penalizing discrimination during optimization, altering user representation to be more neutral, probabilistically modelling the influence of sensitive attributes, or re-ranking unaltered recommendation, all while adhering to different definitions of what discrimination and fairness entails. There is also variation in the application setting of these approaches; most adhere to the regular asymmetric setting where users and items make up fundamentally different concepts, while others consider reciprocal settings where users are recommended to other users like matchmaking. Yet another dynamic is considered in two-sided settings that seek to achieve both consumer- and provider-side fairness concurrently. Despite the great variety, the breadth of consumer-side fairness approaches has yet to be covered in detail by any existing surveys. We further argue for this claim in Sect. 3.3, where we discuss relevant surveys.

In this survey, we have systematically surveyed the existing literature that proposes and evaluates approaches considering consumer-side fairness. Critical aspects of the qualified literature are discussed, compared, and categorized, leading to the proposal of a taxonomy that highlights high-level fairness definitions and how it has been incorporated into the approaches. Further, we provide a comprehensive overview of metrics used to evaluate the fairness of the approaches and some thoughts on the field's future directions. Our key contributions are:

1. Propose a taxonomy for categorizing consumer-side fairness approaches in recommender systems based on high-level conceptual fairness definitions and how the fairness is incorporated.
2. Provide a comprehensive overview, categorization, and comparison of available consumer-side fairness approaches in recommender systems and their proposed fairness evaluation.

The remaining sections of this survey include a background section on fairness definitions, terminology, related concepts, and related works; methodology covering the literature selection process and the proposed taxonomy; a detailed discussion and comparison of the identified literature; analysis of applied fairness metrics and datasets; and a final discussion of our thoughts on the future directions of the topic.

2 Background

As a primer to this survey's core content and discussion, we introduce key established fairness concepts and terms that appear frequently or are subject to ambiguity. The background also covers a discussion of recommender systems concepts related to consumer-side fairness and a look into existing surveys on fairness in recommender systems and how this survey differs.

2.1 Recommender system definition

Recommender systems comprise many varied approaches designed for varied settings and present no single concise definition. This Section introduces recommender systems at a high level and provides a definition designed to aid the discussion of considered methods and evaluation. Readers are encouraged to consult the introduction chapter of the Recommender System Handbook (Ricci et al. 2022b) for general recommender system overviews, motivations and definitions. This survey focuses on personalized recommender systems, i.e., those that seek to accommodate different individuals with customized recommendations based on their individual preferences. The recommender systems discussed will be categorized as either rating-based or ranking-based recommender systems. When applying the notation presented in Table 1, both variations attempt to capture how a set of entities $\mathcal{U}$ will value another set of entities $\mathcal{V}$ on an individual level. $\mathcal{U}$ are typically exemplified as *users* and $\mathcal{V}$ as *items*, and the overall goal is to *recommend* novel items to the users. For rating-based recommender systems, the level objective is to predict individual ratings given by a user u to an item v, r_{uv}, i.e., $\hat{r}_{uv} = r_{uv}$. Ranking-based recommender systems instead take the approach of capturing the general preferences of the users and using this to present the same users with selections of items predicted to be of the users' liking. This selection is ranked, or ordered, according to predicted relevancy, i.e., the topmost item is the one predicted to be the most liked. The resulting objective is analogous to the rating-based objective, $\hat{y}_{uv} = y_{uv}$, but does present slightly different challenges owing to the non-continuous nature of ranking. Both Rating-based and Ranking-based recommender systems may adopt rating data, but Ranking-based recommender systems can more easily adopt data of a more implicit nature, e.g., interaction events.

Table 1 Notation

Symbols	Description
$\mathcal{U}$	Set of all users
$\mathcal{V}$	Set of all items/recommendable entities
$\mathcal{S}$	Set of all possible sensitive attribute configurations
r	Rating in rating-based recommender systems. Double as preference in mixed usage
pref	Intermediate measure of preference used in ranking-based recommender systems
y	Binary indicator for the presence of recommendation in ranking-based recommender systems
$\hat{}$	Modifier, indicate predicted output as opposed to ground truth
${}_{u,v,s}$	Indicate relation with specific users, items or sensitive values respectively
Rec	Set of (top-k) recommendations
Util($\cdot$)	Open-ended/arbitrary utility function
P($\cdot$)	Probability
$\mathbb{1}\{\cdot\}$	Indicator function. Returns 1 when the predicate is true, 0 otherwise
$\hat{\mathbb{E}}[\cdot]$	Arithmetic mean

Recommender systems are implemented using a plethora of different models and methods like Neighbourhood-based Collaborative Filtering (Nikolakopoulos et al. 2022), Matrix Factorization (Koren et al. 2022), various types of Deep Neural Networks (Zhang et al. 2022), Autoencoders (Li and She 2017), Reinforcement Learning (Afsar et al. 2022), Graph-based models (Wang et al. 2021), and various Probabilistic models. Detailed background theories of various models have been left out to avoid significant inflation of the survey's length. However, as this is a comprehensive survey focusing on tangible model proposals, some technical details will be discussed. Readers are encouraged to consult auxiliary sources, like the provided references, when needed.

2.2 Terminology

The following definitions have been added to mitigate confusion stemming from mixing similar terms or different interpretations of specific terms. A low degree of consensus, especially within fairness-focused research, has led to multiple different terms being used for the same concept and other words like *preference* are contextually ambiguous.

Rating In rating-based recommender systems, we are interested in the rating given by a specific user to a specific item and is contrasted with ranking-based recommender systems. Ratings can be discrete and continuous and typically have a set range, e.g., between 1 and 5. Input and target ratings are ratings that are provided by the user and used for training and evaluating the model, while predicted ratings are produced by the model.

Ranking Ordering of items or entities where items near the top are considered more relevant than those below. A ranking outputted by a recommender system is ordered by the modelled preference of the user. Rankings are occasionally explicitly provided by users as inputs or targets but are more frequently a concept applied to evaluate whether the model can identify the most relevant target items, i.e., how high the target items ranked in the predicted ranking.

IR Ranking The field of Information Retrieval comprise an array of different approaches for retrieving information from data storage. We will consider intent the key factor separating IR Ranking and recommender systems: recommender systems seek to suggest novel, but relevant, information to their users while IR Ranking seeks to retrieve the most relevant information. Furthermore, IR Ranking approaches often involve a query and are rarely personalized.

Top-k Top *k* ranked item, where *k* is an integer indicating the number of items that are of interest. *k* is usually quite small, often in the range of 5-20, as user attention is a limiting factor.

Preference (score) Continuous measure of user preference used to produce rankings. Score/value/measure may be omitted in the text if the context allows it.

Ranking-based recommender systems Recommender systems that learn to rank in order to present the user with the top list of suggested items. Typically optimized for predicting if users will *prefer an item over another item*, or just for predicting if users will *like an item.*

Rating-based recommender systems Recommender systems that attempt to match ratings given to items by users, and predict new ratings given by users to unrated items. Typically optimized for *quantifying how much* a user will like an item.

Sensitive attribute Unifying term used to describe demographic attributes that are used to segment users into different groups for which fairness considerations are applied. Other terms such as *demographic* and *private* have been used when the groups are given equal attention, while terms like *protected*, *minority* and *marginalized* are used when one or more groups are of particular concern. *Sensitive attribute* is found to be sufficient for explaining most approaches, but more thorough explanations are provided in cases where asymmetry or special dynamics of a sensitive attribute take a more nuanced role.

Sensitive group A group of users that share a specific instance of a sensitive attribute, e.g., all male users in a setting where gender is considered a sensitive attribute.

Consumer-side fairness Fairness considerations centred on users of recommender systems (Burke et al. 2018). Consumer-side fairness definitions are concerned with *fair* treatment of users on individual or group level. Group-level definitions often group users by sensitive attributes, i.e., Sensitive groups.

Representation Many algorithms commonly used for recommender systems represent users and items as (high dimensional) vectors, e.g., latent factors in factorization models, embeddings in deep models and latent state in autoencoders. In general discussion, we will use *representation* to refer to these concepts.

Prediction performance measure A measure of Prediction Performance evaluates how close predicted recommendations are to target recommendations. Where target recommendations are assumed to reflect user preference, e.g., items rated/liked or interacted with by a user but not known during model training. Most metrics used to evaluate recommender system performance fall under this definition of Prediction Performance measures, e.g., MAE, RMSE, AUC, NDCG, MRR etc. These measures are contrasted by other utility measures discussed in this survey, e.g., statistical parity only considers predictions and not targets. The notion of Prediction Performance measures will play a key role in differentiating between fundamentally different fairness goals.

2.3 Formal fairness definitions

Several formal fairness definitions have been proposed for classification settings, especially those that influence decisions. In these settings, the models are typically tasked with outputting the most likely label or handful of labels given some input. One example could be a bank that utilizes a model to predict whether a potential customer will default should they be given a mortgage. Fairness can be highly relevant in this example, e.g., if the applied model is shown to discriminate through outputting higher probabilities that customers of a specific race or gender will default.

While some fairness definitions targeting such scenarios can be trivially adapted to the recommendation setting, others are more challenging. One such challenge relates to adaptations of definitions that rely on notions like True and False Positives/Negatives, as the interpretations and implications of these statistics may differ between general label prediction and recommender systems. True and False Positives/Negatives are not trivially derived from ratings and are not typically considered in rating-based recommender systems. Conversely, these statistics are heavily influenced by the fixed number of recommendations and the number of correct recommendations in ranking-based recommender systems. Furthermore, the implications of some definitions may be enhanced in scenarios where a positive label is deemed a positive outcome for a stakeholder, even if it was a False Positive. For instance, in our mortgage application example, the applicant will be happy if the application is accepted regardless of whether it was the correct verdict according to bank policies. In consumer-side recommendation settings, this is rarely the case. A False Positive in a top-k recommendation setting will simply be the presence of an item that the user does not care for among the top recommendations.

A selection of fairness definitions is covered here, along with accompanying descriptions of recommender system-specific adaptions. The reader is encouraged to consult Gajane (2017), Caton and Haas (2023), and Li et al. (2023) for a more in-depth discussion of formal fairness definitions in both machine learning and recommender systems. There is also ongoing research into how users perceive different fairness definitions when applied to different scenarios. Harrison et al. (2020) performs a large-scale user study targeting the perceived fairness of machine learning models, while Sonboli et al. (2021) presents a smaller user study specifically targeting recommender systems.

2.3.1 Fairness through unawareness

Fairness Through Unawareness considers any model fair as long as no explicit sensitive attributes are part of the input. This definition is widely disregarded as it fails to consider implicit information on sensitive attributes present in other parts of the input (Gajane 2017).

2.3.2 Statistical parity

Statistical parity for decision problems requires that each group has an equal probability of being assigned a positive outcome, e.g., being granted a mortgage.

$$\mathrm{P}\big(\hat{y} = 1 \mid s = s_1\big) = \mathrm{P}\big(\hat{y} = 1 \mid s = s_2\big) = \ldots$$

Where $\mathrm{P}(\cdot)$ represents probability, $\hat{y}$ is the predicted label, and s is a sensitive attribute.

The recommendation of specific items is rarely considered inherently positive outcomes in recommender systems. Recommender systems also output ratings or rankings instead of predicted labels, which further complicates the adoption of Statistical Parity to evaluate recommender system fairness. A fairness definition for recommender systems inspired by Statistical Parity instead requires that the predicted ratings of items, or the probability of recommending them, is the same for all sensitive groups. This is further discussed in Sect. 3.2.1.

2.3.3 Equal opportunity

Equal opportunity in decision problems requires that the true positive rate of different sensitive groups is equal.

$$P(\hat{y} = 1 \mid y = 1, s = s_1) = P(\hat{y} = 1 \mid y = 1, s = s_2) = ...$$

2.3.4 Equalized Odds

The Equalized Odds definition is stricter than Equal Opportunity in also requiring that the false positive rates of the different sensitive groups are equal.

$$\begin{aligned} P(\hat{y} = 1 \mid y = 1, s = s_1) &= P(\hat{y} = 1 \mid y = 1, s = s_2) = ... \\ \& \, P(\hat{y} = 1 \mid y = 0, s = s_1) &= P(\hat{y} = 1 \mid y = 0, s = s_2) = ... \end{aligned}$$

As previously mentioned, false positives may benefit some stakeholders in certain scenarios. However, false positives may also be the most detrimental type of error in other scenarios. Thus, the decision to pursue Equalized Odds instead of Equal Opportunity may be motivated by a wish to balance either a boon or a bane.

2.4 Research areas related to consumer-side fairness

Many research areas within recommender systems have arisen to cover different needs as they appeared or were made known. It is not uncommon that the concepts considered in different research areas partly overlap, share underlying issues, or share similar mitigation strategies. A number of the most relevant recommender system research areas, when compared with consumer-side fairness, are listed in this section, focusing on similarities and dissimilarities. The intention of this section is two-fold: The first is to highlight related topics that may be of interest to readers and that may help put consumer-side fairness into a broader context. The second motive is to highlight dissimilarities that disqualify certain research from being covered by the scope of this survey, despite occasionally adopting fairness terminology.

2.4.1 Provider-side fairness

As the name entails, provider-side fairness takes a polar opposite view to consumer-side fairness. A significant part of the research focuses on mitigating popularity bias, which occurs when popular items are given disproportional amounts of exposure by the recommender system. However, the broadness of the definition also covers research that is more similar to many consumer-side fairness approaches in considering fairness for groupings based on sensitive information from a provider perspective.

2.4.2 Cold-start, long-tail and diversity

Cold-start, long-tail, and diversity in recommender systems all make out similar concepts with partly overlapping causes and mitigation approaches: **Cold-start** specifically focuses on the scenario of providing good recommendations for new users or new items, through facing the challenge of comparatively little data for the new entity. There are also recommender systems that more generally focus on increasing the recommendation of items in the **Long-tail**, i.e., the items that receive little attention compared to the, typically, smaller amount of popular items. Analogous user-centric recommender systems focus on improving the recommendations for users with few interactions or those receiving sub-par recommendations. Approaches that optimize for **Diversity** attempt to diversify the recommendations given to individual users and the complete user base, motivated by aspects such as popularity bias issues and user experience enhancement. While the definitions differ in generality and perspective, they are sometimes used interchangeably in the literature, especially cold-start and long-tail.

Approaches proposed for addressing issues in these three research areas share many similarities and can directly overlap with approaches proposed to address provider-side fairness issues. A method proposed for mitigating long-tail issues can often also be framed as a method for mitigating popularity bias and ensuring provider-side fairness Item-centric methods for addressing issues related to cold-start, long-tail or diversity share similarities and occasionally overlap, with popularity bias mitigation approaches proposed to improve provider-side fairness. Here, item-centric means the approaches focus on new items, items with few interactions or items that are rarely recommended. Similarly, user-centric approaches that seek to balance the performance of *all individual users* may overlap with consumer-side fairness approaches and are included in this survey, given that they satisfy all acceptance criteria. The last point typically boils down to whether a fairness perspective is applied, i.e., posed as individual fairness, along with fairness evaluation. However, as a general strategy, methods that seek to balance the prediction performance for user groups, where users are grouped based on how many interactions they have, have been excluded. Such approaches fit perfectly within the mature fields of cold-start or long-tail recommender systems and are better represented when compared with such methods.

2.4.3 IR Ranking fairness

IR Ranking is typically not personalized, i.e., the produced rankings are not affected by user-specific interaction with the system. Subsequently, IR Ranking fairness objectives usually have a provider-side point of view, e.g., balancing the exposure achieved by similar items or the representation of different item groups given non-personalized queries. The work of Pitoura et al. (2022) provides an overview of fairness in the IR Ranking setting.

2.4.4 Group recommendation

Group recommendation approaches seek to recommend a set of items to a collection of users, e.g., recommending a travel destination for a group of friends that consider all of their preferences (Masthoff and Delić 2022). Most group recommender systems use aggregated results from individual user models to provide group-level recommendations. How the group-level recommendation respects the individual preferences of the users is central to both the design and evaluation, e.g., one strategy could average the prediction performance while another minimizes the prediction error of the worst-off group member. This research area frequently applies the term *fairness* when explaining their motive of balancing the consideration of the various users in the groups. While this could be accepted as a local fairness definition, i.e., users in each group should be equally satisfied with their recommendation, group recommender systems have been excluded from the survey on the grounds that this is the central mechanic of the research area and setting. Group recommender systems have received significant attention both before and after the advent of fairness research and are better discussed within their research area.

2.4.5 Privacy

Privacy in recommender systems covers many approaches that seek to protect privacy in different stages of the recommender system pipeline. For instance, *federated learning* can be applied to mitigate the issues of having a centralized model that may be breached and mined for sensitive information (Li et al. 2020). *Differential privacy* has been applied to provide protection guarantees for the sensitive information of individual users (Friedman and Schuster 2010). Some privacy approaches seek to completely remove the information of specific attributes within user representations or data, which overlaps with a class of fairness approaches that do the same with the intention of not having the attributes influence the recommendation.

3 Methodology

The methodology of this survey covers the systematic selection process applied for identifying and screening relevant studies, followed by the definition and justification of applied taxonomies, as well as descriptions of how the taxonomies are used to categorize and structure further discussion.

3.1 Selection process

The selection process comprised the definition of concise acceptance criteria, identification of relevant publication indexes, query definition, two rounds of article screenings, and final in-depth reading of the remaining candidate articles. This section presents the acceptance criteria, details the queries and how they were defined, and presents a full overview of the number of studies involved in each step of the selection process.

3.1.1 Acceptance criteria

Five acceptance criteria have been defined in line with our goals of examining the existing literature of tangible models considering consumer-side fairness in recommender systems:

1. The study considers *recommender systems*, see Sect. 2.1.
2. The study considers consumer-side fairness, either explicitly or through a multi-stakeholder focus.

   Note, Group recommendation and Long tail/Cold-start recommender systems are excluded, see Sect. 2.4.
3. The study is published in a peer-reviewed conference or journal.
4. The study proposes a novel model or method.
5. The study evaluates the fairness of the proposed model or method.

3.1.2 Query definition and overview

The keywords were kept general to avoid filtering out potential relevant research. The search queries were chronologically bound by 2016-01-01 and 2022-10-01, where the lower bound was set based on prior knowledge of the topic and preliminary querying for validation purposes. The topic started gaining noticeable traction in 2017, but the early adopters had three publications before this, (Kamishima et al. 2012, 2013, 2016). The first two articles do not appear to have inspired other researchers, but since 2016, there has been a gradual increase in the number of articles each year.

The chronological bound was combined with the keyword combination "recommend*" and "fairness", and both keywords had to be matched in the title, the abstract, or both. "Recommend" was given a wildcard suffix matcher to match both "recommender" and "recommendation". A similar wildcard, "fair*", was used instead of "fairness" in the DBLP index to compensate for not being able to match within the abstracts. Observations in known research and research found through preliminary querying confirmed that all articles that matched "fair" in the title also matched "fairness" when considering the abstract. The wildcard was only used in title-only queries since it significantly increased the number of false positives when matching in both title and abstract. Fairness is becoming a well-established concept within the ML community, and most, if not all, research uses the full term at least once before potentially switching over to the shorthand "fair".

The full selection process is detailed in a PRISMA flow diagram (Page et al. 2021) in Fig. 1.

3.2 Taxonomy

While there have been previous attempts at proposing novel taxonomies for categorizing fairness approaches in recommender systems based on fairness definitions, we argue that there are alternative taxonomies that offer additional insight and value. The most recent taxonomy is proposed by Wang et al. (2022), who first proposes splitting between process and outcome focus, then two alternatives for splitting outcome-focused fairness on target and concept. One challenge when applying this to consumer-side fairness research is that many of the named *concept-based* fairness categories do not occur that often, and the vast majority of identified research would be classified as either optimizing and evaluating for *Process Fairness* or *Consistency Fairness*. We also argue that there may be value in further separating different high-level fairness definitions, e.g., *Consistency Fairness* may consider distance notions that only compare the distribution of predictions given to different groups, but it can also consider distance notions that measure differences in how the predictions of the same groups match the targets. In other words, the former disregards user preference and only focuses on the predicted ratings, while the latter explicitly evaluates prediction performance by matching predicted ratings with target ratings.

We propose a new taxonomy centred on which high-level fairness definition is considered when optimizing and evaluating models. Besides resulting in a balanced segmentation of the identified research, the taxonomy separates key different classes of fairness definitions, some of which fundamentally conflict with each other. For instance, the view that fairness is achieved when the same prediction performance is achieved for all sensitive groups is fundamentally different from the view that each group should be presented with the same recommendations regardless of prediction performance. To

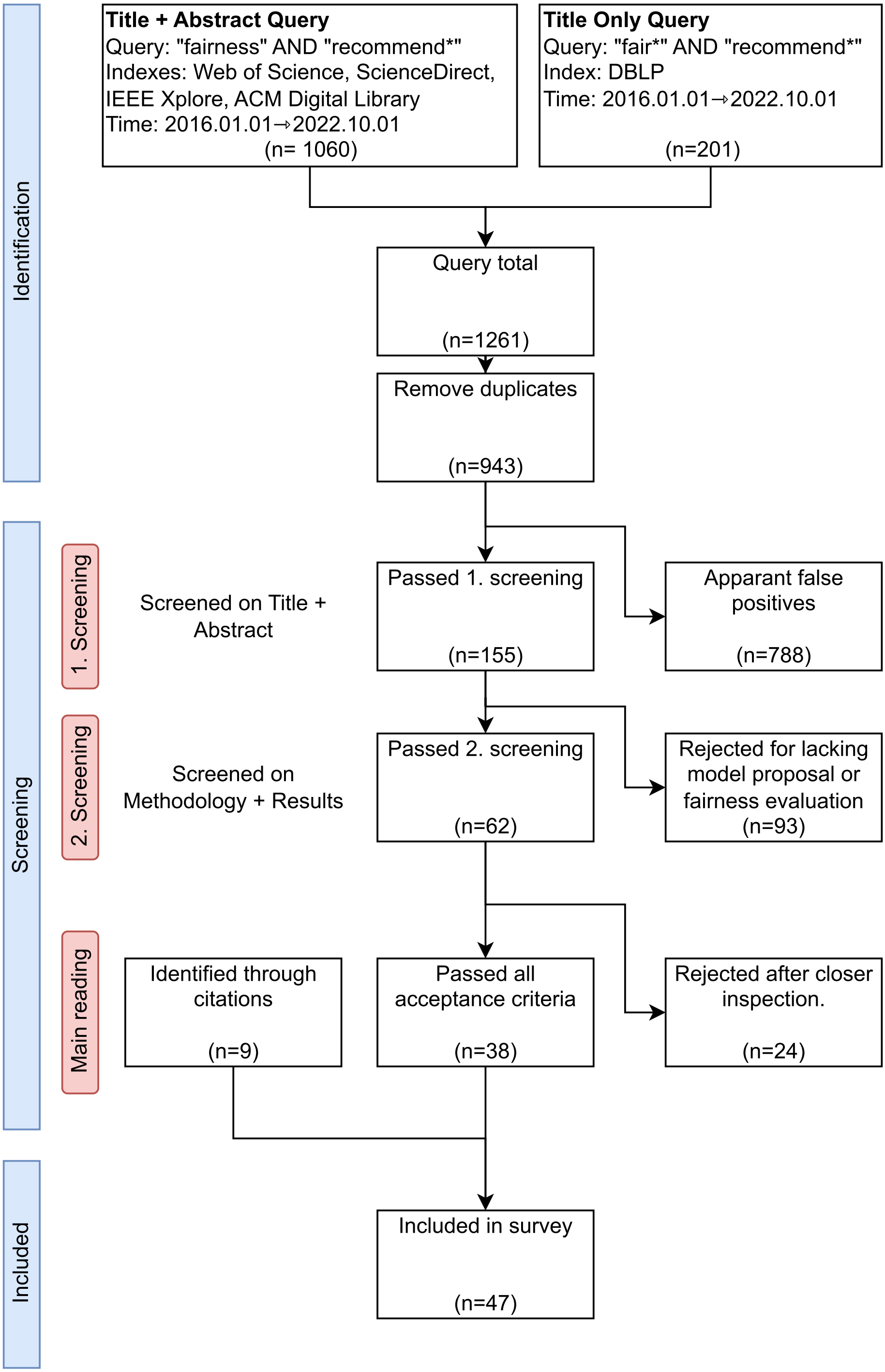
Identification
Title + Abstract Query
Query: "fairness" AND "recommend*"
Indexes: Web of Science, ScienceDirect, IEEE Xplore, ACM Digital Library
Time: 2016.01.01→2022.10.01
(n= 1060)
Title Only Query
Query: "fair*" AND "recommend*"
Index: DBLP
Time: 2016.01.01→2022.10.01
(n=201)
Query total
(n=1261)
Remove duplicates
(n=943)
Screening
1. Screening
Screened on Title + Abstract
Passed 1. screening
(n=155)
Apparant false positives
(n=788)
2. Screening
Screened on Methodology + Results
Passed 2. screening
(n=62)
Rejected for lacking model proposal or fairness evaluation
(n=93)
Main reading
Identified through citations
(n=9)
Passed all acceptance criteria
(n=38)
Rejected after closer inspection.
(n=24)
Included
Included in survey
(n=47)

Fig. 1 A PRISMA flow diagram illustrating the full selection process

further structure and analyze the research, we propose applying two other, more established, concepts which detail how/at which point the fairness consideration is incorporated and which type of recommender algorithm is applied, respectively.

3.2.1 Fairness Interpretation taxonomy

While several fairness definitions from the fields of law and psychology have been formally defined for machine learning, see Sect. 2.3, they cannot trivially be applied for categorizing the studies considered in this survey. The formal definitions are occasionally implemented as metrics, but since they mostly consider the model's outcome, it is challenging to define how they should be adhered to during optimization. Another challenge is that some of these definitions are conceptually similar and only differ in minute details. We instead propose categorizing fairness definitions on a higher and more conceptual level, while remaining compatible with the more low-level formal definitions. For instance, Equality of Opportunity and Equalized Odds share a high-level objective of balancing prediction performance measures evaluated for different sensitive groups. We refer to the high-level fairness definitions as Fairness Interpretations to stress that we are discussing high-level conceptual definitions that cannot be expressed in a single metric but rather express a general idea shared by similar definitions/metrics. Two identified Fairness Interpretations have further been assigned sub-categories for finer distinctions between similar concepts. The taxonomy is illustrated in Fig. 2, and the different Fairness Interpretations are further described in the following sections and illustrated in Fig. 3.

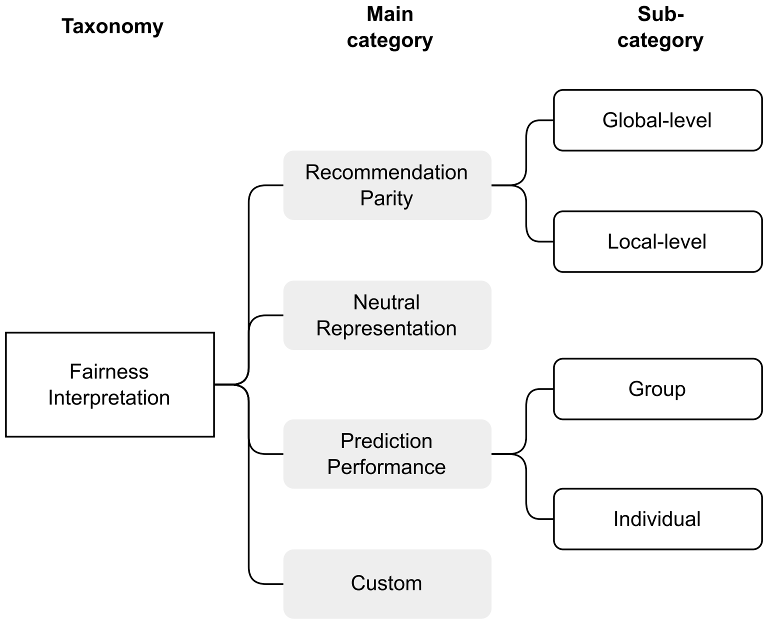

Fig. 2 The proposed taxonomy based on Fairness Interpretation

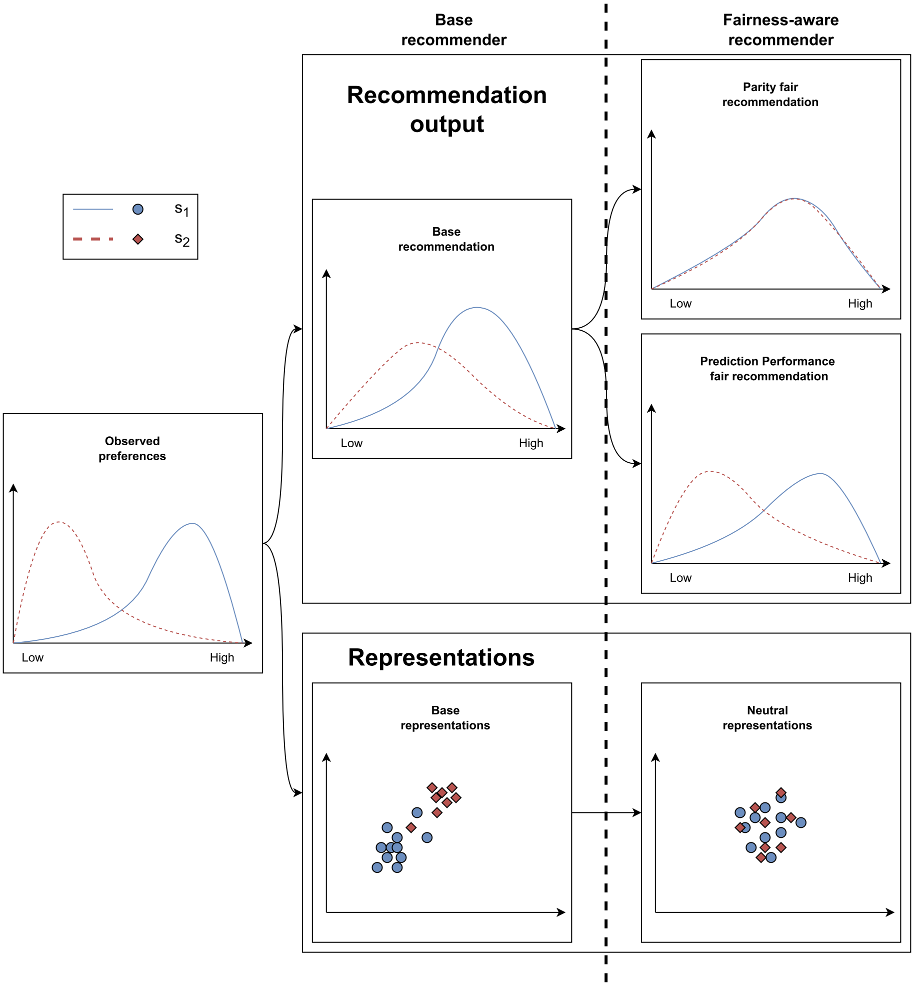

Fig. 3 Diagram that illustrates the high-level differences between three non-Custom Fairness Interpretations in a scenario where the sensitive group s_1 tend to like the items while s_2 dislikes them. In the recommendation and preference plots, frequency is plotted against preference/rating, while representations are projected into two-dimensional scatterplots. In this case, higher prediction performance is achieved for s_1 than for s_2 when the base recommender system is used, e.g., because this group dominates the dataset. Recommendation Parity is satisfied when the recommendation distributions overlap, while Prediction Performance fairness is achieved when the respective recommendation distributions match and mismatch the “true" distributions equally. In the case of representations, the Neutral Representation Fairness Interpretation would seek to move from the case where representations of different groups can be separated into distinct clusters to the case where the clusters overlap or are indistinguishable. **Note** that an example with two sensitive groups was chosen for simplicity. All Fairness Interpretations are well-defined for any number of groups/individuals of two or greater

Recommendation parity

Recommendation Parity considers an approach fair when the probability of recommending an item or the predicted item rating is the same for different sensitive groups. For instance, if age is considered a sensitive attribute, Recommendation Parity requires that senior and young users be presented with the same recommendations. Observed user preference or satisfaction is not considered by this Fairness Interpretation, which contrasts it with Prediction Performance and various Custom Fairness Interpretations.

Recommendation Parity for consumer side fairness can be further split based on at which level the parity is optimized or measured. Some optimize and measure parity at a **global** level, i.e., the distribution of all ratings, while others consider parity in the rating of individual **items** or **item groups**. The former is less constricting since the system can predict that different groups will rate the same item differently as long as the rating differences cancel out globally. For instance, if an item's predicted rating is higher for one group than another, the parity will be regained if there exists another item for which the predicted ratings are mirrored. Local-level Parity demands that the predicted rating or preference of each item/item group is the same for all sensitive groups.

Neutral representation

Neutral Representation considers an approach fair when representations in the model are neutral with respect to sensitive attributes. The key idea is that the sensitive attributes should not influence the recommendation. The research that adopts this Fairness Interpretation varies in having different ideas of what entails influence of sensitive attributes and how neutrality is evaluated, which in turn affects the optimization. For instance, some require no correlation between representation and sensitive attributes, others require representations to be orthogonal to sensitive dimensions in the representation space. Others still define causal models that allow correlation between representation and sensitive attributes but require conditional independence given other variables. The discussion of the methods that are evaluated and optimized for Representation Neutrality is structured by these different ideas.

Unlike the other Fairness Interpretations, the evaluation of Neutral Representation does not consider the outputted recommendation and focuses solely on the model. Achieving perfectly neutral representations will often lead to recommendations akin to those produced by models optimized for Recommendation Parity since the sensitive information no longer affects the predictions. Whether this is the case or not depends on the applied definition of *neutral*. Regardless, given the unique perspective of focusing on representations rather than the recommendations, characteristic optimizations, representation-centric evaluation, and high prevalence of approaches, a dedicated Fairness Interpretation for neutral model representations is still deemed warranted.

Prediction performance fairness

Prediction Performance Fairness considers an approach fair when equal prediction performance is achieved for all individual users or sensitive groups. Unlike Recommendation Parity, this Fairness Interpretation explicitly considers the users' opinion on the recommendations they are provided. Evaluation must be based on user feedback in online settings or target ratings/recommendations in offline settings. Despite this requirement, there is a vast amount of varied fairness definitions and metrics that fall under Prediction Performance Fairness. This is because there are endless ways of measuring how well a prediction matches a target and equally many ways to aggregate or compare these measures between users or sensitive groups.

Custom

The Custom Fairness Interpretation covers fairness definitions that do not fall under the three other Fairness Interpretations. Such fairness definitions have been centred on parity with respect to derived attributes or the balancing of custom utility measures that do not consider prediction performance.

3.2.2 Fairness Incorporation taxonomy

When adapting existing recommender systems to new objectives, e.g., adding fairness objectives, one can classify the adaptations by which part of the recommendation process they target. One such categorization distinguishes between adaptations targeting model input, the model itself and the output. This categorization is well established within the field of recommender system fairness and machine learning as a whole, and the classes are typically named pre-, in- and post-processing, respectively (Caton and Haas 2023; Mehrabi et al. 2021; Deldjoo et al. 2023). To structure the identified research, we propose a taxonomy based on these classes. Each class is extended with an additional level to represent better the observed variety of approaches applied to incorporate fairness awareness. The second level contains a single sub-category each for pre- and post-processing approaches, but we propose four sub-categories to cover the diversity of in-processing methods. The full taxonomy is illustrated in Fig. 4, and each proposed sub-category has been given a brief description in this section.

Data Augmentation

The only sub-category of pre-processing methods is Data Augmentation. Data Augmentation covers all methods that increase fairness by augmenting the model's input data. E.g., modifying the list of movies enjoyed by a male user to look more like the list of a female user when considering Recommendation Parity Fairness.

Loss Enhancement

Loss Enhancement methods improve fairness through additional terms in the loss used for optimizing the model. E.g., adding a loss term that is the mathematical definition of the considered fairness definition. Positive aspects of Loss Enhancement methods are that they can be applied to many recommendation algorithms, are flexible in definition, and can significantly change predictions through minimal changes to an approach. However, extra loss terms do not inherently improve the modelling capacity of a model and may introduce more complex dynamics that would benefit from more modelling capacity or changes to the model architecture.

Probabilistic

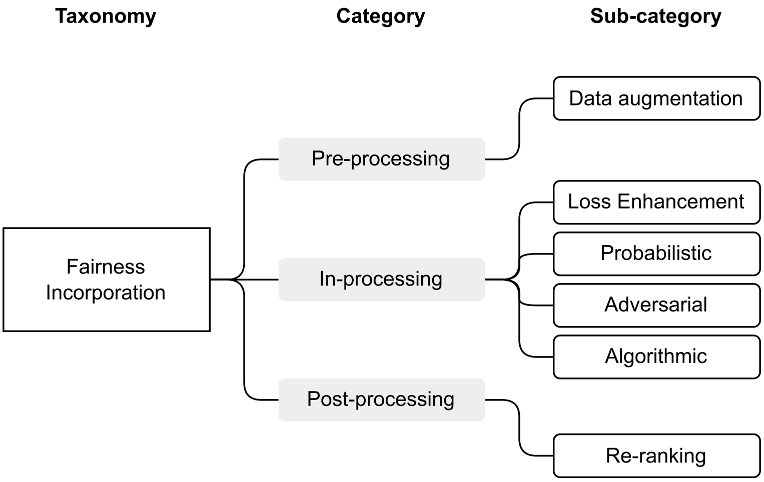

Fig. 4 Fairness Incorporation taxonomy

The probabilistic fairness approaches apply probabilistic concepts to encourage independence of recommendation and sensitive features, apply soft constraints, or filter out sensitive information. Unlike the other in-processing sub-categories, probabilistic fairness approaches are not easily achieved through a smaller extension to an arbitrary model. This variation of Fairness Incorporation usually requires that the applied model is probabilistic in nature itself, at least partially.

Algorithmic

Algorithmic approaches incorporate fairness by changing smaller aspects of an existing algorithm or through one-time processing, e.g., through selective sampling or removal of sensitive projections in representation space.

Adversarial

Adversarial approaches train adversarial models to classify sensitive information encoded in model representations. The adversarial models are then used to inform how the model should be updated to filter out sensitive information better.

Re-ranking

Re-ranking approaches re-rank the recommendation of one or more base recommender systems according to new or changed objectives, e.g., introducing fairness objectives that are optimized along with regular recommendation performance.

3.2.3 Recommendation algorithm

A third categorization is used to classify approaches according to the applied recommender system algorithm. The algorithm can affect how fairness awareness can be incorporated and influence the general recommendation task. Comparing approaches based on the recommendation algorithm is also enhanced by sharing similar implementation details and premises. The Recommendation Algorithm classes have been created by grouping together related recommendation algorithms and are listed by an acronym and a description in the following list.

- *CF* Neighbourhood-based Collaborative filtering. (Nikolakopoulos et al. 2022)
- *MF* Matrix-Factorization. (Koren et al. 2022)
- *NCF* Neural Collaborative Filtering, taken to mean neural network-based collaborative filtering methods that more specialized model groups do not cover. (He et al. 2017b; Zhang et al. 2022)
- *Graph* Various Graph-based models and methods. Graph Neural Networks, Graph Convolutional Networks, Graph Embeddings etc. (Wang et al. 2021)
- *AE* (Variational) Auto Encoders. (Kingma and Welling 2022; Li and She 2017)
- *Probabilistic* Various Probabilistic models and methods. Probabilistic Soft Logic (Kouki et al. 2015), latent models (Govaert and Nadif 2008; Langseth and Nielsen 2012) etc.
- *Classification* Various Classification methods. Logistic Regression, Random Forest (Breiman 2001), Gradient Boosting (Natekin and Knoll 2013) etc.
- *Bandit* Contextual Bandit. (Li et al. 2016)

3.2.4 Structuring of main discussion

The three different categorizations will all be used when discussing and comparing the identified approaches. Three sections are reserved for pre-, in- and post-processing

Fairness Incorporation approaches, and their content is structured by the corresponding Fairness Incorporation sub-categories and both Fairness Interpretation and recommendation algorithms. The Fairness Interpretation taxonomy is used in an overview and for a focused comparison of fairness optimization. In contrast, recommendation algorithms are used to structure a more general technical discussion to highlight comparable implementational choices.

3.3 Related work and taxonomies

There has been a recent surge in proposed surveys of fairness in recommender systems. Pitoura et al. (2022) surveys both fairness in IR ranking and recommender systems, while Deldjoo et al. (2023), Wang et al. (2022), Li et al. (2023) focus on recommender systems. Pitoura et al. (2022) seeks to serve as an overview of fairness in both IR ranking and recommender systems, which makes it the broadest and most high-level survey among the ones considered relevant. They propose using multiple established concepts as a basis for their categorization, e.g., individual/group fairness and provider/consumer side fairness, as well as novel categorizations of fairness incorporation methods. Because of the broad scope and since it is the oldest survey considered relevant, only two of the studies covered in this survey were covered in Pitoura et al. (2022).

Deldjoo et al. (2023), Wang et al. (2022), Li et al. (2023) were all first made publicly available within a few months of each other in 2022, and all consider a broad scope comprising all types of fairness in recommender systems. This scope is wider than the one applied in this survey by covering provider-side fairness and group recommendation. Additionally, all three surveys also cover research that is theoretical in nature or performs analysis using established approaches and datasets, i.e., not necessarily proposing new models or methods. Deldjoo et al. (2023) investigate the current state of fairness in recommender systems, and focus on charting how the available research is distributed with respect to different high-level concepts. They additionally propose a single-level categorization of the fairness metrics considered in the research they cover. Li et al. (2023) has a more conceptual take and provides a thorough introduction to fairness theory and fairness in other machine learning fields before addressing their identified research through multiple different taxonomies based on binary concepts. Some of the fairness concepts they use to categorize the research are well established, like group/individual and consumer/provider, while others have not previously been focused on, like short-term/long-term and black box/explainable. Wang et al. (2022) propose a hierarchical taxonomy of fairness definitions that are since used to categorize the optimization and fairness metrics applied in their identified studies.

Our work differs from previous surveys by specializing in tangible solutions proposed for consumer-side fairness in recommender systems. The specialization allows for a complete overview of the available literature and a higher focus on technical aspects to enhance comparisons. We also categorize our identified research using a new taxonomy centred on high-level fairness definitions and incorporation methods, which has a purposely high-level and general definition to be applicable and extendable to new definitions and methods. The completeness of our survey is exemplified by Table 2, which indicates that when adjusting for time, the largest coverage overlap with the broader surveys only comprises 18 out of the 43 articles we identified in the same time interval.

Table 2 Table displaying the coverage in the most relevant surveys of the articles identified in this survey. Raw counts and percentages are presented, both adjusted and unadjusted for the publish date of the last considered article in each survey

	Adj. coverage	% Adj. coverage (%)	Tot. coverage	% Tot. coverage (%)
Deldjoo et al. (2023)	14/43	33	14/47	30
Li et al. (2023)	18/43	42	18/47	38
Wang et al. (2022)	16/41	39	16/47	34

3.4 Full model overview

This section presents a preliminary analysis and overview of all identified research by recommendation algorithms and the Fairness Incorporation methods. The motive is to put the topic into the broader context of general recommender systems and to provide an overview of all covered research. A full overview is found in Table 3. Note that the same article may fall under multiple types of Fairness Incorporation and recommendation algorithms, since the proposed approach may apply multiple types of Fairness Incorporation strategies and be applied on multiple base models. Also note that even if an incorporation method theoretically can be applied to a recommendation algorithm, only observed combinations are covered. The fact that a method is adaptable for other recommendation algorithms does not guarantee that un-documented combinations will achieve similar results or improvements. Furthermore, the current trends of the field are better reflected when keeping to the combinations that have been actively researched.

3.4.1 Model analysis

Some clear trends can be observed in the full table. The field has experienced rapid growth, with most research taking place in the most recent years. The early adopters focused heavily on pure Loss Enhancement approaches, usually paired with matrix factorization models, but more recent research rarely applies Loss Enhancement as the sole Fairness Incorporation method. Re-ranking methods saw a similar burst of attention in 2020 and 2021 but did not dominate the field, since multiple other directions were researched in the same period. Adversarial approaches were slow to appear but are now actively researched and paired with a varied selection of recommendation algorithms. Bayesian and algorithmic approaches are the smallest in-processing groups but are characterized by being pretty evenly distributed across time and being applied with specific recommendation algorithms. There also appears to be a recent research trend of applying multiple Fairness Incorporation strategies instead of relying solely on a single strategy.

Table 3 All covered research categorized by Fairness Incorporation method and recommendation algorithm

	Pre-processing	In-processing				Post-processing
	Data augmentation	Loss Enhancement	Probabilistic	Algorithmic	Adversarial	Re-ranking
CF	Slokom et al. (2021)	Burke et al. (2018)				Patro et al. (2020b), Ashokan and Haas (2021), Wu et al. (2021c)
MF	Rastegarpanah et al. (2019), Fang et al. (2022), Slokom et al. (2021)	Kamishima et al. (2012, 2013), Kamishima and Akaho (2017), Yao and Huang (2017), Kamishima et al. (2018), Zheng et al. (2018), Wan et al. (2020), Yao and Huang (2021)			Resheff et al. (2019), Li et al. (2021b), Wu et al. (2021b)	Edizel et al. (2020), Patro et al. (2020a), Patro et al. (2020b), Wu et al. (2021c), Ashokan and Haas (2021), Biswas et al. (2021), Do et al. (2021), Wu et al. (2022a)
NCF		Bobadilla et al. (2021), Islam et al. (2021), Wu et al. (2021a), Li et al. (2021a)		Islam et al. (2019), Islam et al. (2021), Li et al. (2022b)	Li et al. (2021b), Wu et al. (2021a, 2022b), Wei and He (2022), Rus et al. (2022)	Wu et al. (2022a)
Graph		Liu et al. (2022c), Liu et al. (2022a)	Buyl and Bie (2020), Li et al. (2022a)	Rahman et al. (2019), Xu et al. (2021), Li et al. (2022b)	Bose and Hamilton (2019), Wu et al. (2021b), Xu et al. (2021), Liu et al. (2022c), Liu et al. (2022b), Liu et al. (2022a)	
Probabilistic			Kamishima et al. (2016), Farnadi et al. (2018), Buyl and Bie (2020), Dickens et al. (2020), Frisch et al. (2021), Li et al. (2022a)			Dickens et al. (2020)
AE		Li et al. (2021a), Borges and Stefanidis (2022)			Borges and Stefanidis (2022)	

Table 3 (continued)

	Pre-processing	In-processing				Post-processing
	Data augmentation	Loss Enhancement	Probabilistic	Algorithmic	Adversarial	Re-ranking
Classification						Paraschakis and Nilsson (2020)
Bandit		Huang et al. (2021)				

4 Pre-processing methods

While numerous studies consider the effect of data augmentation, we only found three papers that pass all acceptance criteria. In particular, several candidates were rejected for not proposing formalized approaches or not presenting an evaluation of the achieved fairness. Pre-processing methods comprise the smallest Fairness Incorporation main category (Table 4).

4.1 Fairness optimization

4.1.1 Prediction performance fairness

Rastegarpanah et al. (2019), Fang et al. (2022) propose exploiting collaborative filtering dynamics by training new synthetic users that will influence the recommendation of the real users. When training synthetic users, Rastegarpanah et al. (2019) enhances the loss by adding terms for penalizing both the group-level and the individual-level variance of rating errors. In contrast, Fang et al. (2022) utilize similar loss terms based on the metrics proposed by Yao and Huang (2017), see Sect. 7.3.1, and also global Recommendation Parity.

4.1.2 Custom

The fairness optimization proposed by Slokom et al. (2021) shares similarities with in-processing approaches optimizing for Neutral Representations but alters the input data to remove correlation between the user profiles and the sensitive attributes of the users, instead of altering the intermediate user representations. The approach achieves this by adding items that are more popular among users of other sensitive groups, identified using auxiliary models, to the user profiles. The authors also explore removing items at random or based on how popular they are in the user's own sensitive group.

4.2 Architecture and method

The three selected papers all propose pre-processing methods that can be applied to a wide variety of recommendation algorithms, and all have used matrix factorization as one of their base recommender models. Rastegarpanah et al. (2019) propose a method for learning supplementary data that can influence polarization and improve individual and group fairness. The key insight is that introducing additional users will affect the recommendation of the original users. This insight is exploited by adding a few synthetic users represented with their own ratings. These ratings are considered parameters and are trained using loss terms designed to influence polarization and fairness. Further, they propose two computationally cheap heuristics. The synthetic data is optimized for prediction performance achieved for both individual users and sensitive groups. Fang et al. (2022) apply the same base approach but focus on optimizing multiple fairness objectives more efficiently and smoothly by projecting the gradients of different objectives onto each other if they conflict. The fairness objectives fall under both Prediction Performance Fairness and Recommendation Parity.

Table 4 Overview of the identified pre-processing approaches structured by the Fairness Interpretation and Fairness Incorporation of their optimization. Approaches that consider multiple Fairness Interpretations are listed in multiple rows

		Data Augmentation
Recommendation Parity	Global	Fang et al. (2022)
	Local	
Neutral Representation		
Prediction Performance	Global	Rastegarpanah et al. (2019); Fang et al. (2022)
	Individual	Rastegarpanah et al. (2019)
Custom		Slokom et al. (2021)

Slokom et al. (2021) modify the data of existing users through an extension of the approach proposed by Weinsberg et al. (2012) instead of training new ones. An auxiliary logistic regression model is trained to tell how indicative items are of the gender of the users that like them. This information is used to select items to be added or removed from user data to make the data less indicative of gender. The addition process specifically intersects lists of indicative items with recommendations from a user-based collaborative filtering model to motivate the addition of relevant items.

5 In-processing methods

In-processing methods are the most represented among the main categories, and their dominance has been constant since the advent of the field. They are characterized by being the most specialized approaches, as the base models themselves are adapted and changed.

5.1 Fairness optimization

5.1.1 Recommendation Parity

Optimization of Recommendation Parity fairness is mainly found among the in-processing methods. Sensitive groups often differ heavily in preferences, so the optimization for a Fairness Interpretation that requires that they are given the same recommendation may benefit from additional influence at a model level and model flexibility. Research that targets Recommendation Parity saw a lot of focus in the field's early years but has since been overtaken by alternatives.

Global Recommendation Parity

Kamishima et al. (2013); Kamishima and Akaho (2017) propose adding loss terms for matching mean rating and preference of different sensitive groups, while Dickens et al. (2020) devise a probabilistic soft logic rule of similar design for the same goal.

More comprehensive approaches for matching global recommendation distributions beyond the first momentum are proposed by Kamishima et al. (2012, 2016, 2018).

Kamishima et al. (2012, 2018) introduce different loss terms for minimizing the mutual information of the ratings and the sensitive groups in matrix factorization. In a slightly different approach, Kamishima et al. (2016) apply a latent factor model where the rating variable is considered independent of the sensitive group variable and optimizes their model using the Expectation Maximization algorithm.

Local Recommendation Parity

In the case of local Recommendation Parity, all relevant research we have found only considers the first moment when matching the recommendations of different sensitive groups. Kamishima et al. (2013) propose adding a loss term that penalizes the squared difference of item ratings between different sensitive groups as an alternative to the already mentioned global version. Similarly, Islam et al. (2021) apply the same idea but opt for an absolute difference instead of a squared difference. The probabilistic soft logic approach proposed in Farnadi et al. (2018) defines rules for encouraging both item-group and item-level parity (Table 5).

5.1.2 Neutral Representation

The objective of Neutral Representation fairness cannot be obtained without altering the model, thus, it is only pursued by in-processing approaches. Optimization of this Fairness Interpretation can be achieved by applying different strategies for filtering out intrinsic sensitive information in representations within the model. The following paragraphs are structured by the technique applied to achieve neutral representations, see also Fig. 4.

Adversarial

The approaches proposed by Resheff et al. (2019), Wu et al. (2021a), Xu et al. (2021), Borges and Stefanidis (2022), Rus et al. (2022) all apply adversarial models directly on model representations. Resheff et al. (2019) pass the latent user factors of their matrix factorization approach to their adversarial, while Wu et al. (2021a) do the same with one of the multiple user representations they train in a composite NCF model. Xu et al. (2021) feed their adversarial model a linear combination of the user representation in a base recommender model and a representation they base on an auxiliary knowledge graph for modelling sensitive user attributes. Rus et al. (2022) propose a neural classification model and applies an adversarial model on a hidden layer in the said model. Finally, Borges and Stefanidis (2022) apply an adversarial model to discriminate the latent representation in their variational autoencoder-based model.

A slightly more intricate scheme is proposed by Wei and He (2022), who concatenate the observed ratings to the representations that are fed to the adversarial model, which they argue will improve the neutrality of the representation and also make the representations independent with respect to the sensitive attribute conditioned on the observed ratings. They further add a second adversarial model, which is fed predicted ratings along with corresponding observed values and item embeddings.

Bose and Hamilton (2019); Li et al. (2021b) argue for letting users dynamically decide which sensitive attributes they are comfortable with the model using. To support this, both propose training optional filters for filtering out different types or combinations of sensitive information from user representations in graph- and matrix-factorization models. The filters are trained using adversarial models. A similar approach is proposed by Wu et al. (2022b), who train *adaptors* (Houlsby et al. 2019) within the *transformers* (Vaswani et al. 2017) that make out their model. The adaptors dynamically

Table 5 Overview of the identified in-processing approaches structured by the Fairness Interpretation and Fairness Incorporation of their optimization. Approaches that consider multiple Fairness Interpretations and Fairness Incorporation methods are listed in multiple rows and columns

		Loss Enhancement	Probabilistic	Algorithmic	Adversarial
Recommendation Parity	Global	Kamishima et al. (2012, 2013); Kamishima and Akaho (2017); Kamishima et al. (2018)	Kamishima et al. (2016); Dickens et al. (2020)		
	Local	Kamishima et al. (2013); Islam et al. (2021)	Farnadi et al. (2018)		

Neutral Representation		Wu et al. (2021a);
		Liu et al. (2022c);
		Liu et al. (2022a)
			Buyl and Bie (2020);
			Frisch et al. (2021);
			Li et al. (2022a)
				Rahman et al. (2019);
				Islam et al. (2019);
				Islam et al. (2021);
				Xu et al. (2021);
				Li et al. (2022b)
					Resheff et al. (2019);
					Wu et al. (2021a);
					Xu et al. (2021);
					Borges and Stefanidis (2022);
					Bose and Hamilton (2019);
					Li et al. (2021b);
					Wu et al. (2021b);
					Liu et al. (2022c);
					Wu et al. (2022b);
					Liu et al. (2022b); Liu et al. (2022a); Wei and He (2022); Rus et al. (2022)

Prediction Performance	Group	Yao and Huang (2017);
		Zheng et al. (2018);
		Wan et al. (2020);
		Huang et al. (2021);
		Yao and Huang (2021);
		Liu et al. (2022c);
		Borges and Stefanidis (2022)
			Dickens et al. (2020)	Li et al. (2021a)	

	Individual				

Custom		Burke et al. (2018);
		Wan et al. (2020);
		Bobadilla et al. (2021)			

filter out different combinations of sensitive attributes based on user- and task-based settings in a sequential recommendation setting.

Wu et al. (2021b), Liu et al. (2022a, 2022b, 2022c) all consider graph neural network methods and the construction of higher-order graph representations by accumulating neighbouring representations in the recommendation graph. The approaches apply adversarial models to discourage the encoding of sensitive attributes in the user- and item-level representations, which also mitigate the accumulation of sensitive information in the higher-order neighbourhood representations. Liu et al. (2022c) further supplements the adversarial discrimination loss with a loss term on the covariance of the predicted attribute and the actual sensitive attribute. Liu et al. (2022a) instead designs and utilizes self-supervising loss terms to enhance the representations and mitigate imbalance issues caused by imbalanced sensitive attributes.

Orthogonality

Orthogonality-based approaches apply additional loss terms or explicit removal of sensitive projections to make representations orthogonal to explicit or implicit sensitive dimensions in the representation space. Wu et al. (2021a) model two separate user representations: one for inferring sensitive information and one for providing neutral representations. They devise a loss term that encourages the two representations to be orthogonal and further encourages the neutrality of the second representation through an adversarial approach.

A more explicit approach is pursued by Islam et al. (2019, 2021), where a post hoc step infers sensitive dimensions in the representation space by taking the difference of the mean representation of each sensitive group. The projections of the sensitive dimension onto each representation are then explicitly subtracted. In the case of Islam et al. (2021), the orthogonality processing supplements the Recommendation Parity loss term (see Sect. 5.1.1).

Sampling based representation training

Rahman et al. (2019); Li et al. (2022b) both adjust the sampling strategy used when training representations. Rahman et al. (2019) proposes to balance the sampling of the next user according to sensitive groups when training graph representations using random walks. In contrast, Li et al. (2022b) adjust the probability of sampling the triplets needed for training knowledge graph representations in a manner that balances out the correlation of sensitive groups and items across all users.

Probabilistic approaches

The models by Buyl and Bie (2020); Li et al. (2022a) are fitted using a prior that is informed of sensitive attributes to allow the rest of the model to focus on other aspects. When the model is used, the sensitive prior is replaced by one oblivious to the sensitive attributes. The intention is to produce fair recommendations along with neutral representations.

Frisch et al. (2021) explicitly model a variable for representing the contribution of the sensitive attributes instead of using a sensitive prior. Ideally, this sensitive variable can soak up all sensitive information, leaving the rest of the model neutral. When recommending, the model drops the parts of the model that are dependent on the sensitive attribute.

5.1.3 Prediction Performance Fairness

When optimizing for Prediction Performance Fairness, models attempt to balance prediction performance measures on a group or individual level. While only group-level optimizations are found among the in-processing methods, there is still significant variation in the considered approaches.

Yao and Huang (2017) proposes four Prediction Performance Fairness metrics for recommender systems, then adapts and applies each metric as loss terms in a matrix factorization approach. One of the metrics is similarly adapted by Dickens et al. (2020) as a probabilistic soft logic rule.

Numerous variations of straight-forward loss terms based on Group Prediction Performance Fairness Interpretations are proposed for different models: the contextual bandit approach proposed by Huang et al. (2021) penalizes differences in cumulative mean rewards of different sensitive groups. Liu et al. (2022c) and Borges and Stefanidis (2022) both supplement adversarial approaches with Prediction Performance Fairness Loss Enhancement. Liu et al. (2022c) penalize the absolute differences in pairwise recommendation loss of different sensitive groups, while Borges and Stefanidis (2022) penalize differences in reconstruction loss of a protected group and that of the average user in their variational autoencoder model. Finally, Yao and Huang (2021) train personalized regularization weights based on the loss of a specific sensitive group to force the matrix factorization model to focus more on the prediction performance achieved for said group.

Wan et al. (2020) consider a unique recommender system setting where users and items are segmented into market segments based on sensitive groups and item groups and argue that the prediction performance achieved for the market segments should be similar. The proposed model applies loss terms that penalize error variation between user groups, item groups, and market segments. The authors also explore a market segment-level parity alternative by penalizing the variances of predicted ratings instead of errors.

Li et al. (2021a) propose a less direct way of encouraging the model to value the prediction performance for non-mainstream user groups more by adding decoder components to their representations and corresponding loss terms for reconstructing the inputs. The intention is to provide the model with a stronger incentive for properly encoding all users and items, which in turn may mitigate issues with favouring the mainstream user groups at the expense of everyone else. A similar goal is pursued by Liu et al. (2022a), who devise a set of auxiliary goals for encouraging their model to produce richer representations of all users.

For the reciprocal setting, Zheng et al. (2018) proposes to consider both the prediction performance for the user that receives the recommendation and the prediction performance for the recommended users themselves. On a global level, this scheme balances the prediction performance for two user groups based on the user's role in individual recommendations, i.e., reference users and users being recommended to reference users.

5.1.4 Custom

Bobadilla et al. (2021) utilize empiric trends in the input data to design a set of indexes that represent users' and items' intrinsic *sensitive value*, e.g., an item is considered fairly young if it is mostly liked by young users, and a user is considered highly senior if they exclusively like item popular with senior users. They further design a loss term to penalize recommending items to users if the index values differ significantly. Loss Enhancement is also applied in the neighbourhood-based collaborative filtering model proposed by Burke et al. (2018) to balance the contribution of peers of different sensitive groups when recommending. Specifically, the added loss term penalizes the absolute difference of the model-specific user-to-user weights of different sensitive groups.

5.2 Architecture and method

5.2.1 Loss Enhancement

Matrix factorization

The early works of Kamishima et al. ; Kamishima et al. are the earliest identified research that satisfies all acceptance criteria in this survey. The four publications (Kamishima et al. 2012, 2013, 2018; Kamishima and Akaho 2017) all propose matrix factorization models where the fairness aspects are modelled through different, but related loss terms. They all share the overall goal and fairness objective of ensuring statistical parity of recommendations. Additionally, all train different sets of parameters for the different sensitive groups they consider. In the first iteration, Kamishima et al. (2012) propose a loss term that is an approximation of the mutual information of the rating and the sensitive attributes. Next, Kamishima et al. (2013) introduces an efficiency improvement with alternative loss terms that penalize differing ratings per sensitive group averaged over all items or individually. The paper by Kamishima and Akaho (2017) considers similar loss terms, but through an implicit-feedback recommender system using ranking-based evaluation. It is noted that the approach has little effect on the ranking order of each user. Finally, Kamishima et al. (2018) returns to rating-based recommender systems, introducing two new methods matching the first and second moment of the distributions to ensure statistical parity. Both methods approximate rating distributions given the sensitive attribute with normal distributions, and then penalize Bhattacharyya distance (Bhattacharyya 1943) and mutual information, respectively.

Another early contribution was by Yao and Huang (2017), who argue for fairness definitions based on balancing prediction performance rather than Recommendation Parity. They propose four new Prediction Performance Fairness metrics that measure imbalances in how well the system recommends for different sensitive groups. Further, they devise loss terms based on these metrics and an additional parity-based metric to compare how well models trained with the different loss terms fare when evaluated using all metrics.

Zheng et al. (2018) is concerned with recommending matches after speed-dating, which is a reciprocal recommendation setting in the sense that consumers are recommended to other consumers. The model predicts user impression of speed-dates with different partners and considers a Custom utility metric based on the similarity of user's expectation of a partner and their impression of the speed-date partner. The utility metric is also used in the added loss term, which is designed to maximize the prediction performance for both users in each potential match, i.e., the user being recommended another user and the recommended user themselves. Considering the opinions of both involved users may also improve the overall success of this specific application, as mutual interest is ideal in a matchmaking scenario.

The approach proposed by Wan et al. (2020) is designed to address retail marketing bias by better balancing the achieved prediction performance for different market segments. In particular, they define market segments based on sensitive user groups and attributes of models used in marketing different items, e.g., one segment may make out male users and items only marketed using female models. The proposed approach optimizes for balanced prediction performance by penalizing error variance between the different segments and other groupings. An alternative configuration is also considered where the model instead penalizes predicted rating variance, i.e., a configuration that optimizes for Recommendation Parity Fairness rather than Prediction Performance Fairness.

The last identified Loss Enhancement-based matrix factorization approach was proposed by Yao and Huang (2021). The key idea of the model is to improve the prediction performance for disadvantaged users through personalized regularization. This is achieved through a multi-step process that alternates between training for the general recommendation task while keeping the personalized regularization weights static and updating the same parameters based on the recommendation loss of the disadvantages.

Neighbourhood-based collaborative filtering

Burke et al. (2018) propose enhancing the loss of a user-based collaborative filtering approach to encourage users' neighborhood of peers to be better balanced with respect the considered sensitive attributes. To this end, they devise a loss term that penalizes if the coefficients used for weighting influence of peers are skewed towards a specific group, i.e., the sum of male peer coefficients is greater than that of female peers.

Neural collaborative filtering

Bobadilla et al. (2021) are unorthodox in terms of fairness definition and approach. They give each item a value based on the distribution of, e.g., the gender of users who like it, thus representing how *gendered* the item is. The items are then used reversely to measure how gendered each *user* is based on the items they like. The authors go on to penalize recommending items with a gendered value far from the user's.

Li et al. (2021a) aims to improve the prediction performance achieved for non-mainstream users using collaborative filtering models. Their approach involves a factorization step that involves user and item representations, where the representations double up as the latent representations in two autoencoders. The autoencoders are added to encourage the model to encode all input information in the latent representations properly, and not neglect information only relevant to a subset of the users.

Bandit

The only identified bandit approach was proposed by Huang et al. (2021), and is a contextual bandit method that penalizes differences in cumulative mean rewards of different sensitive groups. The authors construct a synthetic dataset for video recommendations and define a reward function that, for instance, rewards cases where the gender of the user and the video speaker is the same.

5.2.2 Probabilistic

Graph

Buyl and Bie (2020) consider link prediction in social graphs and applies a probabilistic model for training graph representations based on the work by Kang et al. (2019). They encode prior knowledge of the relations in the network using a prior term, which frees up the representations from encoding the same knowledge. Buyl and Bie (2020) leverage this by designing priors that contain sensitive information to be used during training but replaced in the final recommendations. Li et al. (2022a) further adapt the approach for peer recommendation in online learning and introduce changes to the negative sampling used during training.

Probabilistic model

Farnadi et al. (2018) and Dickens et al. (2020) both apply models based on Probabilistic Soft-Logic (PSL) (Bach et al. 2017) for fairness-aware recommendation. PSL allows probabilistic models using human-interpretable first-order logic. Both models apply a base set of logical rules for the general recommendation task, e.g.,

$$\text{SIMILAR\_USER}(u_1, u_2) \wedge \text{RATING}(u_1, i) \implies \text{RATING}(u_2, i),$$
$$\text{SIMILAR\_ITEM}(i_1, i_2) \wedge \text{RATING}(u, i_1) \implies \text{RATING}(u, i_2).$$

Farnadi et al. (2018) extend the model with fairness rules based on parity, e.g., one sensitive group's rating of an item implies the rating of a different sensitive group and vice versa. Dickens et al. (2020) consider courser parity-based rules and add others to better balance the prediction performance achieved for different sensitive groups. Further, they allow modellers to adjust the importance of different fairness terms. They also discuss using the model together with an arbitrary black-box model to inject fairness and interpretability, which can be thought of as a form of re-ranking.

In the work by Kamishima et al. (2016), two different graphical models are proposed for modelling ratings independent of a sensitive group membership. The models are realized as latent class models and optimized through the Expectation-Maximization algorithm.

Frisch et al. (2021) propose using a latent block model for clustering users and items, then model a new rating-mean based on the associated cluster mean, individual variables for the item- and user-specific influences, and finally an item-specific variable that is controlled by the user's sensitive attribute. This final variable models the sensitive information, and is only used during training, similar to how informed priors are used in Buyl and Bie (2020). The model is optimized using variational inference.

5.2.3 Algorithmic

Neural collaborative filtering

Islam et al. (2019) explicitly subtract sensitive projections in user representations in a neural collaborative filtering model. They consider both scenarios where there is a single or multiple binary sensitive attribute(s), e.g., male/female and young/senior. Some of the same authors (Islam et al. 2021) propose a more complex approach where they utilize transfer learning to pre-train user representations and neural network weights in a non-sensitive recommendation setting. The user representations are then processed similarly as by Islam et al. (2019) to be used in a sensitive recommendation setting. The non-sensitive settings considered are film recommendation and social media action recommendation, while the sensitive settings are occupation and college major recommendation, respectively. A parity-based loss term is applied in addition to the user representation processing to incorporate fairness in sensitive settings.

Li et al. (2022b) propose a fairness-aware sequential recommender system in which an integral part is to train item representations for representing the contextual information of the items and their relations. The authors use fairness-aware sampling when training said representations. Specifically, the sampling probability is set to adjust for any empirical skewness in how an item is preferred by different sensitive groups.

Graph

The approach by Rahman et al. (2019) is designed to be used in reciprocal settings and tested on recommending peers in social networks while considering sensitive groups based on gender and race. The base representation algorithm performs random walks over the graph by sampling the next user among the current user's peers, i.e., the users the current user has a relation to in the observed data. Their fairness view is introduced by first sampling the peer's sensitive attribute uniformly, then sampling as usual from the qualified peers only.

Xu et al. (2021) work with knowledge graph-based recommender systems. They propose training user representations of an auxiliary graph for representing sensitive attributes and their hidden relationships through a multi-layered neural network. This user representation is combined with that of the original recommender system in a linear transformation and then factorized with the item representation from the original recommender system. Additionally, an adversarial network is trained to classify sensitive attributes from the compound user representations and used to filter out said information. The purpose of the auxiliary graph representation is stated to be to improve the modelling of multiple sensitive attributes and their interactions.

5.2.4 Adversarial

Matrix factorization

Resheff et al. (2019) apply an adversarial gradient-based model to remove information like gender and age in the latent user factors. The authors list both privacy and fairness aspects as motivations for adversarial filtering.

Li et al. (2021b) adopt the approach proposed by Bose and Hamilton (2019) using multiple different recommender system specific models, as opposed to the more general setting of link-prediction considered by Bose and Hamilton (2019). The approach is applied using four different models, covering matrix factorization and neural collaborative filtering. They further extend the approach by proposing a secondary option for training single filters for combinations of sensitive attributes, which is compared to the main approach of training one filter for each attribute and taking the mean of filtered representations to apply combinations.

Graph

Bose and Hamilton (2019) proposes to filter combinations of sensitive attributes from graph representations dynamically and considers both link-prediction and recommender system applications using different setups. They train one filter per sensitive attribute and combine filters by aggregating the representations processed by the filters. Each sensitive attribute is further assigned an adversarial for removing the traces of said sensitive attribute. A binary mask is sampled during training to simulate different users who want to restrict the use of different combinations of sensitive attributes. The mask then determines which filters are applied.

Wu et al. (2021b) assume a graph-based perspective and that pre-trained user and item representations are provided. They suggest training filter functions for filtering out sensitive information from both representations and using these to build higher-order neighbourhood representations iteratively. For instance, the first-order neighbourhood representation of a user is based on the filtered representations of the items the user has liked or interacted with, the second-order neighbourhood contains the first-order neighbourhood representation of the same items, and so on. A multi-layered neural network is used to simultaneously process the first- and second-order neighbourhood representations into the final network level representation, with the motivation to capture higher-order sensitive information and how the different levels relate. Adversarial models are applied to both the filtered initial user representations and the final network-level user representations.

Liu et al. (2022a, 2022b, 2022c) also apply neighbourhood representations, along with adversarial models for removing sensitive information in the base representations. However, they differ from Wu et al. (2021b) as they consider end-to-end systems where the base representations are trained as part of the same model, and in using the highest-order representations explicitly as the final representations. The three papers themselves differ in how they construct higher-order neighbourhood representations. Liu et al. (2022a, 2022c) reduce the contribution of higher-order representations by dividing with a function that increases linearly in the order. Liu et al. (2022b) construct higher-order representations by passing the previous-order representations through a neural network and also explicitly considers the representations of neighbours that are increasingly further removed in the graph. The approaches further differ in their application of additional fairness optimization. Liu et al. (2022c) propose two new loss terms: one for penalizing the covariance of the actual sensitive attribute and the one outputted by the adversarial model, and another for penalizing differences in pairwise losses of different sensitive groups. The former further enhances the neutrality of the representations, while the latter penalizes prediction performance imbalance. Liu et al. (2022a) proposes to enhance the base representations by designing and applying a set of loss terms that encourage the representation of more complex information to mitigate the poor representation of underrepresented sensitive groups in the dataset.

Neural collaborative filtering

A neural collaborative filtering model for fairness-aware news recommendation is proposed by Wu et al. (2021a). The key idea is to contain all sensitive information in a part that can be disregarded when the model is applied, similar to the priors in Buyl and Bie (2020). Two separate user representations are trained: one is used for classifying the sensitive attribute and has the intention of aggregating sensitive information, while the other is designed for housing everything else and is coupled with an adversarial model to remove sensitive information. The sum of both user representations is used for recommendation during training while encouraging that the representations are orthogonal through adding a loss term. Only the neutral representation is used once the model is finished training.

Rus et al. (2022) propose to improve fairness in a job recommender system by filtering out gender information in representations of user resumes. To this end, they first train new word embeddings on resumes and job application texts within a proprietary dataset. The trained word embeddings are then used to encode the resume texts, and the encoded texts serve as inputs to their proposed neural recommender model. An adversarial model is applied to filter out gender information at a specific layer in the multi-layered neural network model. The authors also explore a simple alternative where they instead replace gendered words with neutral words before training the word embeddings. However, the adversarial approach is shown to outperform this alternative.

Wu et al. (2022b) follow Bose and Hamilton (2019); Li et al. (2021b) in letting the users decide which sensitive attributes can be considered when generating recommendations. However, while the preceding research train multiple filters for different sensitive attributes that are dynamically plugged into the recommender system pipeline when activated, the filtering components in Wu et al. (2022b) are static parts of the model that dynamically change behaviour based on personalized prompts concatenated with the input. The filtering components are based on the *adaptor* proposed by Houlsby et al. (2019) and trained along with different discriminators while keeping the remaining model parameters frozen.

The framework proposed by Wei and He (2022) considers multiple Fairness Interpretations simultaneously. The framework consists of two main loops: an inner loop in which users are initialized with the latest parameters suggested by a meta-model and then optimized for different tasks, and an outer loop where the results of the inner loop are used to update the parameters of the meta-model to produce better user initializations in the next cycle. The framework applies two different adversarial models, which both attempt to detect sensitive attributes: the first one is fed user representations based on trained context representations and the users' observed ratings, while the second considers the predicted ratings, the corresponding observed ratings, and item representations.

Variational autoencoder

Borges and Stefanidis (2022) use a variational autoencoder (VAE) as their main model. The VAE is considered a collaborative recommender, where the decoded latent representation is interpreted as an encoding from which recommendations are extracted. The VAE is extended with an adversarial model for training the model to produce neutral latent representations and a loss term for encouraging the model to be equally good at reconstructing the inputs of a specific *protected* sensitive group as at reconstructing the inputs of all users on average.

6 Post-processing methods

Post-processing methods share one of the main benefits of pre-processing methods in being flexible with respect to which recommender system model is used. Additionally, post-processing methods do not affect the raw data but are arguably the least flexible approaches since they are constrained by the provided recommendation and the data used to train the model. Post-processing methods are not as popular as in-processing methods but have received more attention than pre-processing methods (Table 6).

6.1 Fairness optimization

6.1.1 Recommendation Parity

Both the post-processing techniques for Recommendation Parity included in the survey use the *global* perspective. Ashokan and Haas (2021) propose using results from the training data to align the predicted ratings of the two sensitive groups better in a binary setting. To this end, they add the mean rating difference of the two sensitive groups during training to the predicted ratings of one of the groups when using the model for a new recommendation.

The approach in Dickens et al. (2020) discussed in Sect. 5.1.1 is also applicable as a re-ranker of a base recommender model. Thus, their proposed probabilistic soft logic rule comprises a second identified strategy for optimizing Global Recommendation Parity in post-processing methods.

6.1.2 Prediction Performance Fairness

Group

In addition to the Recommendation Parity optimization of Ashokan and Haas (2021) covered in Sect. 6.1.1, the authors also propose a similar optimization targeting Prediction Performance Fairness. Here, the per-item average rating error for each sensitive group, as observed in the training data, is added to the individual ratings of the users. The intention is to correct for expected errors observed for different sensitive groups.

The only identified approach that optimizes group Prediction Performance Fairness in a two-sided fairness approach is proposed by Wu et al. (2022a). Through applying the

Table 6 Overview of the identified post-processing approaches structured by the Fairness Interpretation and Fairness Incorporation of their optimization. Approaches that consider multiple Fairness Interpretations are listed in multiple rows

		Re-ranking
Recommendation Parity	Global	Dickens et al. (2020)
	Local	Ashokan and Haas (2021)
Neutral Representation		
Prediction Performance	Group	Ashokan and Haas (2021); Wu et al. (2022a)
	Individual	Wu et al. (2021c)
Custom		Edizel et al. (2020); Paraschakis and Nilsson (2020); Patro et al. (2020a); Patro et al. (2020b); Biswas et al. (2021); Do et al. (2021)

Frank-Wolfe algorithm (Frank and Wolfe 1956), the authors optimize for consumer-side fairness by minimizing the variance of a differentiable definition of NDCG, along with the general recommendation and provider-side fairness objectives.

Individual

A multi-stakeholder approach is proposed by Wu et al. (2021c), in which the consumer-side fairness objective is to fairly distribute the loss of prediction performance incurred by the producer-side exposure considerations among the users. They devise a two-step approach, where the first step identifies highly preferred items that still have to reach their maximum exposure in the recommendation lists of the users. These items then have their ranks fixed, i.e., they will have the same ranks in the final recommendation. Each user is assigned one item at a time in a manner to even out the benefit of being processed first. The second step fills in the free recommendation slots with the items that still require exposure per the provider-side objective.

6.1.3 Custom

Patro et al. (2020a, 2020b); Biswas et al. (2021) all consider multi-stakeholder settings where the consumer-side objective is to distribute the loss of utility among users fairly. Unlike Wu et al. (2021c), they propose applying a utility definition that is not tied to the ground truth, i.e., the considered utility does not measure prediction performance. Their shared utility measure is purely based on the preference values outputted by the original recommender system and produces values indicating how far from *optimal* the new recommendations are deemed. Patro et al. (2020a) and Biswas et al. (2021) both propose similar two-step approaches as that of Wu et al. (2021c) but opt for guaranteeing the producer-side objective in the first step by allocating items in need of exposure in turn to users while attempting to prioritize items preferred by the user. The second step fills the remaining slots with the users' most preferred items. Finally, multi-sided fairness in a dynamic setting is explored by Patro et al. (2020b) who attempts to retain both provider- and consumer-side fairness when facing different incremental changes to the recommendation, e.g., a gradual transition to a new base recommender model. Individual fairness is preserved by introducing lower-bound user utility constraints in the proposed integer linear programming model.

A similar setting is considered by Do et al. (2021), whose approach also optimizes for two-sided fairness and applies custom user utility tied to the base recommender's outputted preference scores system. The approach positions itself to maximize the custom utility of worse-off users and items simultaneously in both regular and reciprocal recommendation settings.

Edizel et al. (2020) focus on providing users recommendations that are uncorrelated with their sensitive attributes. The goal shares many parallels with in-processing approaches that optimize for intermediate representations that are uncorrelated with the sensitive attributes but operate on sets of recommendations due to the post-processing nature. The key idea is to allow users to inherit the recommendations of similar or arbitrary users belonging to different sensitive groups, thus muddying the correlation between the sensitive attributes and both recommendation lists and individual recommendations.

Another re-ranking approach is by Paraschakis and Nilsson (2020), which considers an individual fairness definition based on calibration by user preferences. Specifically, they consider a matchmaking setting where users specify how important it is for them to date within their race or religion. The problem is optimized through dynamic programming.

6.2 Architecture and method

6.2.1 Neighbourhood-based collaborative filtering

The approach proposed by Ashokan and Haas (2021) is technically a re-*rating* approach since they consider rating-based recommender systems but share many similarities with re-ranking approaches. The key idea is to attempt correct predictions of members of different sensitive groups by adding the average prediction errors for each item and sensitive group, as observed in the training set, to new predictions. A second parity-based option instead adds the average difference of the predicted ratings given to different sensitive groups for each item.

6.2.2 Matrix factorization

Edizel et al. (2020) focus on ensuring the generated top-*k* recommendations are uncorrelated with sensitive groups by making users inherit the recommendations of other users belonging to a different sensitive group. While parts of the approach are enhanced when the same top-*k* recommendations are recommended to multiple users by the base recommender system, the approach also works when all top-*k* recommendations are unique, which is not unlikely for a large *k* and a large item catalogue. Two different schemes for recommendation inheritance are evaluated: random and based on similarity. The former is shown to be more effective at reducing how indicative the recommended set is of the user's sensitive group but also impacts prediction performance more negatively.

The works produced by Patro et al. (2020a, 2020b), Biswas et al. (2021) have an overlapping set of authors, and all consider the same two-sided recommendation setting. In Patro et al. (2020a), the primary goal is to satisfy minimum requirements for provider-side exposure, subject to a secondary goal of distributing the loss of a custom utility measure among the users fairly. They propose a two-step approach where users are first allocated one item in turn until each item has achieved its minimum exposure criteria. The users are given the best remaining item according to said user's predicted preferences. Secondly, the remaining recommendation list of each user is filled based on the original recommendation. The approach is proven to guarantee recommendations that are *envy-free up to one item* (EF1) (Budish 2011), i.e., no user would prefer the recommendations given to other users, given that one good recommendation is removed from the recommendations. The subsequent work by Biswas et al. (2021) improves the model by identifying and removing *envy-circles*, which are directed cycles in a graph representation of user envy, without affecting the EF1 guarantee. Patro et al. (2020b) look into incrementally updating the recommendation according to new data or major changes while retaining user and provider-side fairness. They consider three scenarios: switching the underlying recommender system, incorporating new data, and considering additional features in relevance estimation. For this approach, user-side fairness is the primary goal, and its performance is enforced through constraints in the proposed integer linear programming approach.

Similarly, Do et al. (2021) also optimizes consumer- and provider-side fairness with respect to custom utility measures based on the base recommender's predictions and ranking exposure. Their key insight is to consider the utility of each user and item an objective, but ordering the objectives by performance for Pareto efficiency comparisons, e.g., the utility of users u_1 and u_3 are compared if they achieve the worst utilities in the two compared solutions. This objective formulation and ordering render Pareto efficiency equivalent to Lorenz efficiency, meaning that the utility of the worse-off users and items is maximized, and the model is optimized using the Frank–Wolfe algorithm.

Wu et al. (2021c) considers a similar two-sided setting as the work above but opens for providers having more than one item each. Further, the provider exposure is not considered uniform across recommendation ranks but higher in better positions on the recommended lists. For each recommendation position, the approach iterates through the users in an order based on the current recommendation utility and attempts to fill the position with a high-ranking item in the user's original recommendation, subject to maximum restrictions on provider exposure. Unfilled positions are since filled with items for meeting exposure requirements.

Another two-sided approach is found in Wu et al. (2022a), which differs from the aforementioned two-sided approaches by relying more heavily on established optimization models. The consumer-side fairness objective is set to minimize the variance of smooth, i.e., differentiable, NDCG. In contrast, the provider-side fairness objective similarly considers the variance of exposure achieved by different items. The setup is applied by considering the base recommendation of multiple models, including two matrix factorization models and one neural collaborative filtering model, and produces multiple Pareto optimal solutions. The final solution is selected among these by favouring the solution that yields the best outcome for the worst-off objective. Thus, unlike Patro et al. (2020a); Wu et al. (2021c) whose methods prioritized provide- and consumer-side fairness, respectively, through the order in which the objectives were considered, this approach does not explicitly favour one objective.

6.2.3 Classification methods

A return to matchmaking recommender systems is found in Paraschakis and Nilsson (2020), which details a calibration-based fairness approach. The approach considers the case where users define their preference for dating within their race or religion on a percentage scale, and the fairness objective is defined as making sure that the recommended list of each user has a racial or religious composition close to their preference.

One proposed way to model the fair recommendation problem is to frame it as a Knapsack problem and find optimal solutions using dynamic programming. They also propose an alternative Tabu-search-based approach, which scales better but does not guarantee an optimal solution. The fairness evaluation is based on the same calibration applied throughout the approach as a fairness definition.

7 Metrics

A plethora of different metrics has been proposed and applied in the identified research. Contributing factors to this great number of metrics are varying fairness definitions and the fact that different recommender system settings pose different requirements, e.g., rating vs ranking, binary vs multivalent sensitive attributes, and user-item vs user-user recommender systems. Another contributing factor is that the topic only recently gained popularity, and there is a subsequent lack of consensus on evaluation. The metrics have all been structured according to the Fairness Interpretation they fall under, and liberties have been taken in grouping similar metrics under new descriptive names. Further, formulas have been adapted and rewritten in the same notation for consistency. A lookup table for said notation can be found in Table 1. Each subsection covers a Fairness Interpretation and presents a table of the identified metrics, key contexts, and a list of the research that applied them. Sensitive attributes that are not binary, i.e., they can take more than two values, are referred to as **Multivalent**. The table column **Rec. Setting** covers whether a recommender system considers the classical case with designated users and items, user-item, or the reciprocal setting where *users* take the role of both user and item, user-user.

Readers should note that the vast majority of the metrics measure a notion of discrimination or a lack of fairness. In other words, the fairness objective is usually to minimize the metrics. For the metrics where this is not the case or is not clear, explicit descriptions are provided.

7.1 Recommendation Parity metrics

7.1.1 Global-level Parity

The Global-level Parity was first used as a metric by Yao and Huang (2017) and is simply the absolute difference of the mean predicted rating or preference score of different sensitive groups. The metric has been used to evaluate many of the considered approaches but is usually supplemented with more intricate metrics (Table 7).

$$\&\#Xarrowvert;\hat{\mathbb{E}}_{v\in\mathcal{V}}\left[\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_1}}\left[\hat{r}_{uv}\right]\right]-\hat{\mathbb{E}}_{v\in\mathcal{V}}\left[\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_2}}\left[\hat{r}_{uv}\right]\right]\&\#Xarrowvert; \tag{1}$$

7.1.2 Item-level Parity

Three identified metrics share a general design for aggregating the disparity of ratings or recommendations at the item level. All three metrics measure the item-level difference of ratings/recommendations aggregated by sensitive groups and a final aggregation by items.

$$\hat{\mathbb{E}}_{v\in\mathcal{V}}\left[\,\shortmid\,\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_1}}[\hat{r}_{uv}] - \hat{\mathbb{E}}_{u\in\mathcal{U}_{s_2}}[\hat{r}_{uv}]\,\shortmid\,\right] \tag{2}$$

Bose and Hamilton (2019); Wu et al. (2021b) apply identical metrics that consider the simple absolute difference of item ratings for different groups in binary-sensitive attribute settings. Bose and Hamilton (2019) also consider multivalent sensitive attributes, for which Eq. 2 is expanded to consider all possible pairs of sensitive groups.

$$\hat{\mathbb{E}}_{v\in\mathcal{V}}\left[\&\#Xarrowvert;\ln\left(\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_1}}\left[1\{\hat{y}_{uv}\}\right]\right) - \ln\left(\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_2}}\left[1\{\hat{y}_{uv}\}\right]\right)\&\#Xarrowvert;\right] \tag{3}$$

$$\max_{v1,v2\in\mathcal{V}\shortmid v1\neq v2}\&\#Xarrowvert;\left(\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_1}}\left[1\{\hat{r}_{uv_1} > \hat{r}_{uv_2}\}\right]\right) - \left(\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_2}}\left[1\{\hat{r}_{uv_1} > \hat{r}_{uv_2}\}\right]\right)\&\#Xarrowvert; \tag{4}$$

Islam et al. (2021); Frisch et al. (2021) both define similar concepts named ϵ-(differentially) fair, where individual ϵ's reflect how much the recommendation of a single item differs in a binary sensitive group setting. The former considers the probability of recommending items to different sensitive groups, Eq. 3, while the latter considers the probability of ranking an item higher than another item to different sensitive groups, Eq. 4. Islam et al. (2021) takes inspiration from *differential privacy* (Dwork 2011) and adopts logarithmic terms in the absolute difference. However, differential privacy considers a maximum absolute difference, unlike the authors who compute the average ϵ. Frisch et al. (2021) does not cite differential metrics or concepts but is concerned with the maximum ϵ and not the average. Out of the two aggregations, the maximum poses a stronger guarantee than the average and is more in line with the differential definition.

7.1.3 Item-level Rating Deviation

While the Item-level Parity metrics consider the mean difference of sensitive groups' predicted item ratings, Xu et al. (2021) opt for measuring the mean standard deviation of the same ratings. Standard deviation is inherently capable of considering more than two aggregated ratings, and its squared differences make it penalize large differences more. This metric may be a good choice when there are many sensitive groups, and it is important that no group have predicted ratings far from those of other groups.

$$\begin{gathered}\hat{\mathbb{E}}_{v\in\mathcal{V}}\sqrt{\hat{\mathbb{E}}_{j\in\mathcal{S}}\left[\left(\hat{\mathbb{E}}_{u\in\mathcal{U}_j}[\hat{r}_{uv}] - \mu_v\right)^2\right]},\\ \text{where } \mu_v = \hat{\mathbb{E}}_{k\in\mathcal{S}}\left[\hat{\mathbb{E}}_{u\in\mathcal{U}_k}[\hat{r}_{uv}]\right].\end{gathered} \tag{5}$$

7.1.4 Mutual information, rating

Mutual information is a concept from information theory, comprising a measure of the mutual dependency of two variables. Kamishima et al. (2012) proposed to apply mutual information as a fairness metric for recommender systems by measuring the mutual dependency of the ratings and the sensitive attributes. The mutual information is zero if the ratings and the sensitive attributes are independent, i.e., the recommendations given to all sensitive groups are the same. Measuring true mutual information is often intractable, and research has approximated the measure in different ways, e.g., empiric approximations

Table 7 Overview of identified Recommendation Parity metrics, their key properties and the research that has applied them

Name	Fairness subcategory	Sensitive attribute	Rec. setting	Rec. type	Research
Global-level Parity	Global Parity	Binary	User-Item	Rating	Yao and Huang (2017) Farnadi et al. (2018) Dickens et al. (2020) Ashokan and Haas (2021) Fang et al. (2022) Rus et al. (2022)
Item-level Parity	Local Parity	Binary[a]	User-Item	Mixed	Bose and Hamilton (2019) Islam et al. (2021) Frisch et al. (2021) Wu et al. (2021b)
Item-level Rating Deviation	Local Parity	Multivalent	User-Item	Ranking	Xu et al. (2021)
Mutual Information, Rating	Global Parity	Binary	User-Item	Rating	Kamishima et al. (2012) Kamishima et al. (2013)
Kolmogorov-Smirnov	Global Parity	Binary	User-Item	Mixed	Kamishima et al. (2016) Kamishima and Akaho (2017) Kamishima et al. (2018)
χ^2-test	Local Parity	Binary	User-item	Ranking	Frisch et al. (2021)
Group-to-group Variance	Global Parity	Multivalent	User-User	Ranking	Rahman et al. (2019) Buyl and Bie (2020)
Sensitive-group Share	Global Parity	Multivalent	User-User	Ranking	Rahman et al. (2019)

[a] Binary definition has also been applied to multivalent settings by applying it for all possible pairs

of probabilities and bucketing the ratings into intervals to replace the inner integral with a sum.

$$\sum_{s \in \mathcal{S}} \int \mathrm{P}(\hat{r}, s) \log \frac{\mathrm{P}(\hat{r} \mid s)}{\mathrm{P}(\hat{r})} d\hat{r} \tag{6}$$

7.1.5 Kolmogorov–Smirnov statistic

The Kolmogorov–Smirnov(KS) statistic is used to measure how different two probability distributions are and is defined as the greatest difference between the cumulative distributions. Kamishima et al. (2016) proposed using the metric to measure parity by approximating cumulative rating distribution using predicted ratings for different sensitive groups.

$$\sup_{\hat{r}} \mid F_{s_1}(\hat{r}) - F_{s_2}(\hat{r}) \mid \tag{7}$$

F are cumulative distributions, and sup is the supremum, meaning that the statistic returns the upper-bounded difference between cumulative distributions of ratings. Perfect fairness is achieved if the statistic is zero.

7.1.6 χ^2-test

χ^2-tests are typically used to determine if the difference in collections of categorical data is probable, given that they were sampled from the same distribution. Frisch et al. (2021) applied χ^2-test to test the independence of clusters of users and user gender, where the user clusters are defined by the proposed Latent Block Model. Cluster membership influences the predicted rating in the model, so achieving clusters with similar gender compositions as the overall gender compositions will lead to predicting similar ratings for different genders. The test assigns a probability of the user group and gender being independent, which is maximized when optimizing for Recommendation Parity.

7.1.7 Group-to-group variance

In the reciprocal setting considered by Rahman et al. (2019); Buyl and Bie (2020), each recommendation involves two users who may belong to different sensitive groups. Rahman et al. (2019) proposed the Group-to-group Variance metric to measure the variance of the rate at which each pair of sensitive groups is recommended. For instance, given two sensitive groups and that A-B is interpreted as the rate at which users of sensitive group A are recommended users of sensitive group B: the metric uses the rates A-A, A-B, B-A and B-B to calculate the rate variance.

$$\begin{aligned} &\hat{\mathbb{E}}_{a,b \in \mathcal{S} \times \mathcal{S} \mid a \neq b}[(N_a - N_b)^2], \\ &\text{where } N_a = N_{s_i,s_j} = \frac{\mid \{\hat{y}_{u_1,u_2} \mid u_1 \in \mathcal{U}_{s_i}, u_2 \in \mathcal{U}_{s_j}\} \mid}{\mid \mathcal{U}_{s_i} \times \mathcal{U}_{s_j} \mid}. \end{aligned} \tag{8}$$

7.1.8 Sensitive-group Share

Similar to the Group-to-group Variance metric, Rahman et al. (2019) proposes another metric for the reciprocal recommender system setting. The Sensitive-group Share metric measures how well an individual sensitive group s is represented in the recommendations of all users in a reciprocal setting. It subtracts the real representation ratio for said group from the ideal uniform ratio, such that the output represents how far from ideal the recommendations are with respect to single sensitive groups. A single sensitive group's fairness is optimized when the absolute value of the metric is minimized, and the metric itself approaches zero.

$$\mathcal{S}\text{-share}(s) = \frac{1}{|\mathcal{S}|} - \frac{\sum_{u_1 \in \mathcal{U}} \sum_{u_2 \in \mathcal{U}_s} \frac{1\{u_2 \in \text{Rec}_{u_1}\}}{|\text{Rec}_{u_1}|}}{|\mathcal{U}|} \tag{9}$$

7.2 Neutral Representation metrics

Neutral Representation differs from Recommendation Parity and Prediction Performance Fairness in that it does not explicitly concern itself with the actual outputs of the model. This also extends to the metrics of the Fairness Interpretation (Table 8).

7.2.1 Sensitive Reclassification

The vast majority of research optimizing for sensitive neutral representations performs some form of evaluation of how well sensitive information can be extracted from model representations. This evaluation is usually performed by training an auxiliary classification model to classify sensitive information given representations. The classification score becomes an inverse measure of how well sensitive information has been eliminated. *Accuracy*, *F1 score* and *Area Under the ROC Curve*(AUC) are all metrics that have been used for this purpose. AUC is the total area under the curve of the curve you get by plotting the True Positive rate and the False Positive rate while moving the threshold used to split positive and negative classifications. AUC is by far the most applied classification metric.

Given perfectly neutral representations, no classifier will achieve better scores than that of a random classifier, meaning that this metric is typically minimized towards the performance of a random classifier. For instance, the optimal AUC score given a binary sensitive attribute is 0.5.

7.3 Prediction Performance Fairness

Prediction Performance Fairness metrics all consider the balancing of some measure of prediction performance achieved for individual users or sensitive groups. The measures of prediction performance can be based on user feedback in online evaluation settings and measures of how well ratings/rankings match the targets in offline evaluation settings. Most metrics used to evaluate recommender system performance qualify as such measures. The

Table 8 Overview of identified Neutral Representation metrics, their key properties and the research that has applied them

Name	Sensitive groups	Rec. setting	Rec. type	Research
Sensitive Reclassification	Multivalent	Mixed	Mixed	Bose and Hamilton (2019) Resheff et al. (2019) Buyl and Bie (2020) Li et al. (2021b) Wu et al. (2021b) Li et al. (2022a) Borges and Stefanidis (2022) Wu et al. (2022b) Wei and He (2022)

metrics found in this section are often flexible in being able to consider different prediction performance measures but differ in how the prediction performance measures are compared, e.g., absolute difference, standard deviation etc (Table 9).

7.3.1 Group Rating Error Difference

Yao and Huang (2017) propose four metrics that have been adopted in the evaluation of multiple models considered in this survey.

$$\begin{aligned}&\text{ValueUnfairness}\\&=\hat{\mathbb{E}}_{v\in\mathcal{V}}\left[\left|\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_1}}[\hat{r}_{uv}-r_{uv}]-\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_2}}[\hat{r}_{uv}-r_{uv}]\right|\right]\end{aligned}\tag{10}$$

$$\begin{aligned}&\text{AbsoluteUnfairness}\\&=\hat{\mathbb{E}}_{v\in\mathcal{V}}\left[\left|\left|\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_1}}[\hat{r}_{uv}-r_{uv}]\right|-\left|\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_2}}[\hat{r}_{uv}-r_{uv}]\right|\right|\right]\end{aligned}\tag{11}$$

$$\begin{aligned}&\text{OverestimationUnfairness}\\&=\hat{\mathbb{E}}_{v\in\mathcal{V}}\left[\left|\max\left(0,\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_1}}[\hat{r}_{uv}-r_{uv}]\right)-\max\left(0,\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_2}}[\hat{r}_{uv}-r_{uv}]\right)\right|\right]\end{aligned}\tag{12}$$

$$\begin{aligned}&\text{UnderestimationUnfairness}\\&=\hat{\mathbb{E}}_{v\in\mathcal{V}}\left[\left|\max\left(0,\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_1}}[r_{uv}-\hat{r}_{uv}]\right)-\max\left(0,\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_2}}[r_{uv}-\hat{r}_{uv}]\right)\right|\right]\end{aligned}\tag{13}$$

Each metric accumulates absolute differences in per-item errors of two sensitive groups, and all of them involve mechanics for cancelling errors in case the model is similarly underperforming for both groups. The latter two focus on over- and under-estimation, respectively, while the former two consider both error types concurrently while differing in how the errors can cancel out. Specifically, Value Unfairness, Eq. 10, allows errors of the same type to cancel out, while Absolute Unfairness, Eq. 11, allows all errors to cancel out regardless of error type.

Table 9 Overview of identified Prediction Performance Fairness metrics, their key properties and the research that has applied them

Name	Fairness subcategory	Sensitive groups	Rec. setting	Rec. type	Research
Group Rating Error Difference	Group	Binary[a]	User-item	Mixed	Yao and Huang (2017) Farnadi et al. (2018) Dickens et al. (2020) Islam et al. (2021) Wu et al. (2021b) Fang et al. (2022)
Group Rating Error Deviation	Group	Multivalent	User-item	Mixed	Xu et al. (2021)
Group Performance Difference	Group	Binary[a]	Mixed	Mixed	Zheng et al. (2018) Islam et al. (2019) Huang et al. (2021) Liu et al. (2022c) Borges and Stefanidis (2022) Liu et al. (2022b) Liu et al. (2022a) Wei and He (2022)
Performance Variance	Mixed	Multivalent	User-item	Mixed	Rastegarpanah et al. (2019) Wu et al. (2021c) Wu et al. (2022a)
Performance Delta	Group	Multivalent	User-item	Mixed	Li et al. (2021a) Slokom et al. (2021)
Mutual Information, Relevance	Group	Binary	User-Item	Ranking	Kamishima and Akaho (2017)
Inequality of Odds	Group	Binary	User-item	Ranking	Li et al. (2022a)
Inequality of Opportunity	Group	Binary	User-item	Ranking	Kamishima and Akaho (2017)
Protected Performance	Group	Binary	User-item	Rating	Yao and Huang (2021)
GEI	Group	Binary	User-item	Rating	Ashokan and Haas (2021)
Generalized Cross Entropy	Group	Multivalent	User-item	Ranking	Liu et al. (2022c)
F-Statistic Performance	Group	Multivalent	User-item	Rating	Wan et al. (2020)

[a] Binary definition, but has also been applied to multivalent settings by taking the mean or standard deviation of all possible pairs/one-vs-all values

$$\hat{\mathbb{E}}_{v\in\mathcal{V}}\left[\left|\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_1}}\left[\left|\hat{r}_{uv}-r_{uv}\right|\right]-\hat{\mathbb{E}}_{u\in\mathcal{U}_{s_2}}\left[\left|\hat{r}_{uv}-r_{uv}\right|\right]\right|\right] \tag{14}$$

A metric proposed by Wu et al. (2021b), Eq. 14, resembles Absolute Unfairness but considers the absolute errors at the user level instead of at the group level. This incurs a higher penalty when the errors of the predicted ratings vary a lot within one or both groups, as they are not evened out by taking the group-wise mean before the absolute error is calculated.

7.3.2 Group Rating Error Deviation

Along with Item-level Rating Deviation, Xu et al. (2021) propose another metric that considers the standard deviation of item-level statistics of different sensitive groups. This time, it is a Prediction Performance Fairness metric, with a central term considering the squared difference of mean user-level absolute item rating errors. While the metric is structurally similar to Group Rating Error, see in particular Eq. 14, it is important to note that this metric measures the mean *standard deviation* of the group-wise rating errors instead of the mean *difference* of the same errors. Further, the Group Rating Error Deviation metric is inherently capable of considering multivalent sensitive attributes and will incur larger penalties when the achieved prediction performance for user groups is significantly better or worse than that achieved for other groups.

$$\begin{gathered}\hat{\mathbb{E}}_{v\in\mathcal{V}}\left[\sqrt{\hat{\mathbb{E}}_{i\in\mathcal{S}}\left[\left(\hat{\mathbb{E}}_{u\in\mathcal{U}_i}\left[\left|\hat{r}_{uv}-r_{uv}\right|\right]-\mu_v\right)^2\right]}\right],\\ \text{where } \mu_v=\hat{\mathbb{E}}_{j\in\mathcal{S}}\left[\left(\hat{\mathbb{E}}_{u\in\mathcal{U}_j}\left[\left|\hat{r}_{uv}-r_{uv}\right|\right]\right)\right].\end{gathered} \tag{15}$$

7.3.3 Group Performance Difference

Among the considered research, some explicitly calculate the absolute difference of Recall, Precision and NDCG (Järvelin and Kekäläinen 2002) achieved by different sensitive groups. Also included in this group of metrics are more implicit comparisons of prediction performance through tables or graphs. MAE and various custom metrics have been compared through such means. A special case of this class of metrics is applied in the reciprocal setting of Zheng et al. (2018), where the users receiving the recommendation comprise one sensitive group, and the users that make up the recommended entities comprise the other sensitive group.

$$\left|\,\mathrm{Util}(\mathcal{U}_{s_1})-\mathrm{Util}(\mathcal{U}_{s_2})\,\right| \tag{16}$$

7.3.4 Performance variance

The variation of the prediction performance achieved for individual users or sensitive groups has been applied as fairness metrics and was first proposed by Rastegarpanah et al. (2019). It is particularly suited for representing the prediction performance spread when there are too many scores to cover individually, i.e., many users or sensitive groups. The identified variations of Performance Variance have centred around the metrics Mean Squared Error and NDCG.

$$\hat{\mathbb{E}}_{u_1 \in \mathcal{U}}\left[\hat{\mathbb{E}}_{u_2 \in \mathcal{U}, u_2 \neq u_1}\left[(\mathrm{Util}(u_1) - \mathrm{Util}(u_2))^2\right]\right] \tag{17}$$

$$\hat{\mathbb{E}}_{s_1 \in \mathcal{S}}\left[\hat{\mathbb{E}}_{s_2 \in \mathcal{S}, s_2 \neq s_1}\left[(\mathrm{Util}(\mathcal{U}_{s_1}) - \mathrm{Util}(\mathcal{U}_{s_2}))^2\right]\right] \tag{18}$$

7.3.5 Performance Delta

The Performance Delta metrics consider how the prediction performance achieved for specific sensitive groups changes when a fairness-aware model is compared to baselines and were first considered by Li et al. (2021a). From a Prediction Performance Fairness view, a decrease in prediction performance achieved for a dominant sensitive group may be worth a subsequent increase in the prediction performance achieved for worse-off sensitive groups. Having measures of how the prediction performance for specific user groups has changed is useful when considering such trade-offs. Slokom et al. (2021) also consider the absolute difference of the Δ-s of two sensitive groups to capture the dissymmetry in the magnitude of the changes, Eq. 20.

$$\Delta_s = \mathrm{Util}(\mathcal{U}_s) - \mathrm{Util}_{\mathrm{baseline}}(\mathcal{U}_s) \tag{19}$$

$$\Delta_{\mathrm{diff}} = \left| \Delta_{s_1} - \Delta_{s_2} \right| \tag{20}$$

7.3.6 Mutual information, relevance

Kamishima and Akaho (2017) applies mutual information in a Prediction Performance Fairness evaluation of their ranking-based recommender system. The definition is structurally identical to that of the Recommendation Parity metric in Sect. 7.1.4. The main difference is that the predicted rating variable is replaced with a binary relevancy variable, i.e., the degree of independence between relevant recommendations and the sensitive attribute is measured.

7.3.7 Inequality of Odds

Equality of Odds requires the true positive rate (TPR) and false positive rate (FPR) of different sensitive groups to be equal. Li et al. (2022a) propose a metric based on this definition for measuring the level of discrimination. The same research also applies Absolute Between-ROC Area(ABROCA), proposed in Gardner et al. (2019), which inherently considers all possible positive-boundary thresholds as opposed to a single fixed threshold.

$$\max\left(\left| \mathrm{FPR}_{s_1} - \mathrm{FPR}_{s_2} \right|, \left| \mathrm{TPR}_{s_1} - \mathrm{TPR}_{s_2} \right| \right) \tag{21}$$

7.3.8 Inequality of Opportunity

Kamishima and Akaho (2017) considers a metric based on Equality of Opportunity, which requires equality of true positive rates (TPR) of different sensitive groups. They opt for measuring discrimination with a non-absolute difference to avoid abstracting away the orientation of imbalances.

$$\mathrm{TPR}_{s_1} - \mathrm{TPR}_{s_2} \tag{22}$$

7.3.9 Protected Performance

The Protected Performance metrics only consider the prediction performance achieved for specific sensitive groups that are considered protected and in need of improved prediction performance. Yao and Huang (2021) is the only identified research that adopts this metric class, and the authors propose using RMSE to quantify prediction performance. Fair models can be optimized to minimize or maximize this metric depending on the considered prediction performance metric.

7.3.10 Generalized Entropy Index

Generalized Entropy Index(GEI) is a metric often used for measuring income inequality and has been applied both for measuring individual inequality and inequality among larger user groups in recommender systems. In the case of GEI based on individuals, s is interpreted as each user having their own sensitive group $\mathcal{U}_s$. Note that the applied utility measure typically will average over all represented users to get comparative scores when calculating the utility of all users and the one based on a subset of the users. Farnadi et al. (2018) consider GEI for both $\alpha = 1$ and $\alpha = 2$, where the former yields a special case known as the Theil index. All variations measure inequality and have to be minimized to improve fairness.

$$\mathrm{GEI}(\alpha) = \begin{cases} \frac{1}{|\mathcal{S}|\alpha(1-\alpha)} \sum_{s\in\mathcal{S}} \left(\left(\frac{\mathrm{Util}(\mathcal{U}_s)}{\mathrm{Util}(\mathcal{U})} \right)^{\alpha} - 1 \right), & \alpha \neq 0, 1 \\ \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \frac{\mathrm{Util}(\mathcal{U}_s)}{\mathrm{Util}(\mathcal{U})} \ln \frac{\mathrm{Util}(\mathcal{U}_s)}{\mathrm{Util}(\mathcal{U})}, & \alpha = 1 \\ -\frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \ln \frac{\mathrm{Util}(\mathcal{U}_s)}{\mathrm{Util}(\mathcal{U})}, & \alpha = 0 \end{cases} \tag{23}$$

7.3.11 Generalized Cross Entropy

Generalized Cross Entropy allows the evaluator to adjust the importance given to the prediction performance achieved for different sensitive groups through a custom probability distribution and a hyperparameter named β. The metric was first proposed to evaluate recommender system fairness by Deldjoo et al. (2021), and is particularly useful in settings where some users are considered premium users, or there are known marginalized users for whom prediction performance should be improved.

$$\mathrm{GCE}(m,S)=\frac{1}{\beta\cdot(1-\beta)}\left(\sum_{s\in S}\mathrm{P}_f^{\beta}(s)\cdot\mathrm{P}_m^{(1-\beta)}(s)-1\right)\tag{24}$$

P_f is a fair probability distribution, e.g., a uniform distribution if the motive is to give equal focus to all sensitive groups. P_m is a probability distribution over the model m's utility, i.e., the measure of prediction performance. The choice of utility measure determines whether the metric should be minimized or optimized to improve fairness.

7.3.12 *F*-statistic performance

The F-Statistic is typically calculated as part of a F-test used in analyses of variance. Wan et al. (2020) applies the F-statistic while considering the variance of rating errors between market segments and within them as a measure of fairness, where lower F-statistics indicate higher levels of fairness.

7.4 Custom fairness

Custom fairness encompasses all definitions and measures that do not strictly consider Recommendation Parity, Neutral Representation or Prediction Performance Fairness. Unlike Prediction Performance Fairness metrics, any utility considered in these metrics is not affected by how well the predictions match user preference. Additionally, while some of these metrics consider parity-based concepts, the parity relates to aspects other than recommendation, i.e., not explicitly rating, preference or ranking. For instance, parity with respect to derived/contextual user attributes or a particular item group (Table 10).

7.4.1 Normalized Ranking Change

Patro et al. (2020a, 2020b); Biswas et al. (2021) all consider a two-sided fairness setting and interpret the consumer-side fairness as how far the recommendations deviate from the original ranking when producer-side fairness has been taken into account. I.e., they use the intermediate preference scores of the re-ranked top-k recommendations normalized by the optimal preference scores of the original top-k recommendations as a proxy of the recommender system performance. This measure's mean and standard deviation are considered in their fairness evaluation. When optimizing for this fairness definition, the mean ranking change is maximized towards one, while the standard deviation is minimized towards zero.

$$\mathrm{PrefUtil}_u(\mathrm{Rec})=\frac{\sum_{v\in\mathrm{Rec}}\mathrm{pref}_{uv}}{\sum_{v'\in\mathrm{Rec}_{\mathrm{org\_u}}}\mathrm{pref}_{uv'}}\tag{25}$$

$$\mathrm{NormRankChgMean}=\hat{\mathbb{E}}_{u\in\mathcal{U}}\left[\mathrm{PrefUtil}_u(\mathrm{Rec}_u)\right]\tag{26}$$

$$\mathrm{NormRankChgStd}=\sqrt{\hat{\mathbb{E}}_{u_1\in\mathcal{U}}\left[\hat{\mathbb{E}}_{u_2\neq u_1}\left[\left(\mathrm{PrefUtil}_{u_1}(\mathrm{Rec}_{u_1})-\mathrm{PrefUtil}_{u_2}(\mathrm{Rec}_{u_2})\right)^2\right]\right]}\tag{27}$$

Table 10 Overview of identified Custom fairness metrics, their key properties and the research that has applied them

	Sensitive groups	Rec. setting	Rec. type	Research
Normalized Ranking Change	Multivalent	User-item	Ranking	Patro et al. (2020a) Patro et al. (2020b) Biswas et al. (2021)
Ranking Change Envy	Multivalent	User-item	Ranking	Patro et al. (2020a) Biswas et al. (2021)
Gini Coefficient, Preference	Multivalent	Mixed	Ranking	Do et al. (2021)
Protected Item-group Recommendation Parity	Binary	User-item	Ranking	Burke et al. (2018)
Preferential Calibration	Multivalent	User-item	Ranking	Paraschakis and Nilsson (2020)
Intrinsic Sensitive Attribute Match	Binary	User-item	Ranking	Bobadilla et al. (2021)
Sensitive Neutral Recommended Items	Multivalent	User-item	Ranking	Li et al. (2022b)
Segment Recommendation Frequency Parity	Multivalent	User-item	Ranking	Wan et al. (2020)
Sensitive Reclassification, Pre-/Post-	Multivalent	User-item	Mixed	Edizel et al. (2020) Wu et al. (2021a) Slokom et al. (2021)

7.4.2 Ranking Change Envy

In multi-sided fairness settings, it is not uncommon to have a situation where the prediction performance achieved for one user may be increased by giving said users the recommendations given to another user instead of their own. The term *envy* has been taken to refer to cases where users would fare better given the recommendations of other users. Envy can arise in multi-sided recommendendation fairness when auxiliary considerations affect some users more than others, e.g., considerations of provider-side fairness. Patro et al. (2020a) and Biswas et al. (2021) consider the aggregated envy of their proposed *preference utility*, Eq. 25. Fairness is optimized when Rank Change Envy is minimized.

$$\begin{aligned} \text{RankingEnvy}(u, u') &= \max(\text{PrefUtil}_u(\text{Rec}_{u'}) - \text{PrefUtil}_u(\text{Rec}_u), 0) \\ \text{RankChgEnvy} &= \hat{\mathbb{E}}_{u_1 \in \mathcal{U}}\left[\hat{\mathbb{E}}_{u_2 \neq u_1}\left[\text{RankingEnvy}(u_1, u_2)\right]\right] \end{aligned} \tag{28}$$

7.4.3 Gini coefficient, preference

Do et al. (2021) propose adapting the Gini coefficient, frequently used to measure wealth inequality, to measure individual consumer fairness in their two-sided fairness setting. The Gini coefficient is measured for a custom utility measure based on the outputted preference scores of their base recommender system and the ranking positions of the re-ranked recommendations. Their proposed utility measure is defined in Eq. 29, where $\boldsymbol{P}_{\boldsymbol{uv}}$ is a row vector of probabilities for recommending user u item v in different ranking positions, and $\boldsymbol{w}$ is a column vector of exposure/rank weights for the same ranking positions. The metric considers pair-wise differences of the custom utility measure, meaning that lower values represent less inequality.

$$\text{PrefUtilExp}(u) = \sum_{v \in \mathcal{V}} \text{pref}_{uv} \boldsymbol{P}_{\boldsymbol{uv}} \boldsymbol{w} \tag{29}$$

$$\text{GiniPref} = \frac{\sum_{u_1, u_2 \in \mathcal{U} \times \mathcal{U}} | \text{PrefUtilExp}(u_1) - \text{PrefUtilExp}(u_2) |}{2 \, |\mathcal{U}|^2 \, \mathbb{E}_{u \in \mathcal{U}}[\text{PrefUtilExp}(u)]} \tag{30}$$

7.4.4 Protected Item-group Recommendation Parity

Burke et al. (2018) applied this metric to evaluate how balanced the recommendation of the protected item groups are between two sensitive groups and considers a film dataset and a micro-loan dataset. The protected item groups were selected among film genres that are unevenly recommended to different genders for the film dataset, and among the most unfunded regions in the micro-loan dataset.

$$\frac{\sum_{u \in \mathcal{U}_{s_1}} \sum_{v \in \text{Rec}_u} \frac{\gamma(v)}{|\mathcal{U}_{s_1}|}}{\sum_{u' \in \mathcal{U}_{s_2}} \sum_{v' \in \text{Rec}_{u'}} \frac{\gamma(v')}{|\mathcal{U}_{s_2}|}} \tag{31}$$

γ is a function that returns "1" if the item belongs to any protected item groups and "0" if none of its item groups is protected. The ideal metric value is 1, indicating that protected item groups are recommended equally to the different sensitive groups.

7.4.5 Preferential Calibration

Paraschakis and Nilsson (2020) consider a matchmaking scenario where users can set their preference for being matched with people belonging to the same sensitive group as them. The proposed metric measures how well the provided recommendation respects the user's preferences, a concept known as calibrated recommendation (Steck 2018).

$$1 - \frac{\delta_u - \delta_{\text{min\_}u}}{\delta_{\text{max\_}u} - \delta_{\text{min\_}u}} \tag{32}$$

First, the optimal recommendation composition is calculated based on the provided user preference and the sensitive group composition of the full population. δ_u is then calculated as the absolute difference between the ideal and actual composition. Finally, the normalized δ is subtracted from 1 after identifying the best and worst possible δ-s for the actual user, yielding a metric for which values closer to 1 indicate that the users' preferences are better respected.

7.4.6 Intrinsic sensitive attribute match

Bobadilla et al. (2021) devise a notion of intrinsic sensitive properties in both items and users. They assign values to the said property for items by first considering the ratio of female users who like the items compared to the ratio of female users who dislike them, then taking the difference between that value and the equivalent value calculated for male users. These item values are then used reversely to assign values to the same intrinsic properties of the individual users. The fairness of the individual recommendations is set to be the squared difference between the intrinsic user value and the intrinsic item value.

$$\text{IntrinsicSensitiveMatch}(u, v) = (\text{UserIntrinsic}_u - \text{ItemIntrinsic}_v)^2 \tag{33}$$

7.4.7 Sensitive Neutral Recommended Items

Here $\text{SensitiveEntropy}_v$ is the information entropy of the sensitive attribute of the users involved in the interactions with the item v. The information entropy of an item is maximized when different sensitive groups historically have interacted with it at an identical rate, which is considered ideal by the fairness definition considered by Li et al. (2022b). The full metric is the difference between the summed entropy of the recommended items and the summed entropy of the *ground truth* recommendations, which can be interpreted as how much more neutral are the recommended items compared with the *correct* recommendations, i.e., improved fairness will increase this metric.

$$\begin{aligned} \text{IF}(u) &= \sum_{v \in \text{Rec}_u} \text{SensitiveEntropy}_v \\ \text{DIF}(u) &= \text{IF}(u) - \text{IF}_{\text{ground\_truth}}(u). \end{aligned} \tag{34}$$

7.4.8 Segment Recommendation Frequency Parity

Wan et al. (2020) considers a segmented market, where each segment covers a group of users and a group of products. Within this setting, they argue that the distribution of recommendations given across segments should match the observed data across the same segments. To that end, they construct frequency distributions that represent how the segments are represented in the observations and the recommendations and calculate a distance, i.e., discrimination, between the two using *Kullback–Leibler* divergence *of* the recommended frequencies *from* the observed frequencies. This distance is non-negative and zero given perfect parity.

7.4.9 Sensitive Reclassification, Pre-/Post-

Analogous to the Neutral Representation metric Sensitive Reclassification, Sect. 7.2.1, some studies measure how well sensitive attributes can be identified using the input or output data. The pre-processing approach of Slokom et al. (2021) reports the AUC achieved by auxiliary classifiers tasked with identifying sensitive attributes given user data modified by their approach. Similarly, the in-processing approach of Wu et al. (2021a), and the post-processing approach of Edizel et al. (2020) collectively report accuracy, macro-F1 and *Balanced Error Rate*(BER) achieved by analogous classifiers that are fed recommendation sets. BER is the sum of the False Positive rate and the False Negative rate divided by two.

8 Datasets

Like all machine learning models, recommender systems rely heavily on the datasets used to train them, i.e., the recommender systems are optimized to capture the correlations that can be observed in the datasets. The datasets are also pivotal in evaluating and comparing different approaches, and can highlight how well the approaches perform, scale, and generalize. While there is no lack of actors that could benefit from recommender systems and who possess vast amounts of user data to train models on, the sensitive nature of user data often limits the viability, or even legality, of sharing the data publicly. The sensitive nature of the data is further enhanced if sensitive attributes of the users are included, which is required when training many fairness-aware approaches. Furthermore, high-profile examples demonstrating that anonymization of data may not suffice in protecting the identity of the users (Narayanan and Shmatikov 2008), along with an increasing focus on user protection in international discourse and legislation, have likely further deterred actors from sharing their data.

There is a conspicuous lack of relevant datasets for evaluating consumer-side fairness in recommender systems. This discrepancy is both in terms of the total number of available datasets and the relevancy of the domains they represent. The ethical ramifications of discriminatory recommender systems are better highlighted by a career recommendation setting than through a movie recommendation setting. While it is not unlikely that learned social roles partly explain current differences in the movie preferences of male and female consumers, the further propagation of such preferences is arguably less severe than consistently recommending low-income careers to female career seekers because of biases in the data.

8.1 Dataset overview

An overview of all datasets that contain information on sensitive attributes can be found in Table 11 and covers the application domain, the presence of a selection of consumer-side sensitive attributes and a tally of the number of studies that have evaluated their models on the specific datasets. The MovieLens datasets(Harper and Konstan 2015) dominate the field. Among the variations of the dataset, the 1-million and the 100-thousand alone contain sensitive consumer attributes in the form of gender, age and occupation. A wide adoption of the same dataset poses numerous benefits, like improved comparisons and calibrations. However, one ideally wants multiple widely adapted datasets, as different datasets usually pose different challenges, and good performance on one dataset does not necessitate good general performance. Eight studies considers various datasets based on LastFM, who all share domain but vary in size, scope and time of collection. The second most adapted *singular* dataset applied for consumer-side fairness is the Google Local Review dataset(He et al. 2017a), yet it is only considered by a total of four different studies, neither of which consider the MovieLens dataset. Of the remaining datasets, only a handful are used for evaluation in more than a single study, and many of these only appear more than once because they are applied in multiple studies by the same research group. For instance, three of the four studies using the Google Local Review datasets share a subset of authors. It is safe to say that the field currently lacks an established set of benchmark datasets.

Regarding the domains covered by the different datasets, most cover the recommendation of digital media, commodities or services. The few datasets that present more sensitive scenarios have not managed to attract attention in a field that is starved for relevant data, which may imply other limiting factors like restricted availability, dataset size and data quality. In particular, the aforementioned privacy concerns likely play an essential role in the lack of relevant datasets.

When factoring in dataset occurrence counts, most studies consider datasets that provide information on the consumer's gender, age and occupation. Among these, gender is the most widely adopted sensitive attribute and is typically split into two groups, male and female. The adoption of age as a sensitive attribute is also prevalent, and the attribute has been split into two or more groups based on age intervals. Occupation is rarely used, which has been attributed to difficulties related to the high number of highly skewed labels that make empiric evaluation difficult and possibly misleading.

When considering one or more sensitive attributes of a user base, one will typically conclude that the dataset is imbalanced with respect to some configurations of said attributes. Each additional attribute will split each existing user group into two or more smaller groups that may exhibit imbalance. For instance, the typical user in the MovieLens dataset is a young male, while more than 75% of the users in IJCAI2015 are female. Such imbalance can pose significant challenges to fairness goals as models optimized for prediction performance often will end up favouring the typical users. The choice of dataset strongly influences how challenging it is to alleviate the bias stemming from the data, and achieving good fairness in spite of high levels of imbalance can be a testament to a robust method.

The datasets listed in Table 12 do not explicitly provide sensitive information and have either been used to evaluate individual fairness or have supplemented such information using other means. For instance, Bose and Hamilton (2019) compiled a dataset using the Reddit API (Reddit 2022), only comprising users who have explicitly provided gender in communities they partake in that require this. Paraschakis and Nilsson (2020) consider matchmaking and use demographic data on the religious adherence of different races in the US to

Table 11 Table representing the different datasets applied in fair consumer-side recommender systems research with sensitive attributes

	References	Setting	Sensitive attribute							Count			
			Gender	Age	Marital status	Occupation	Race/ Nation- ality	Region	Body size	Pre-	In-	Post-	Total
MovieLens	Harper and Konstan (2015)	Films	✓	✓		✓				2	22	3	27
LastFM[a]	Last.fm (2022)	Music	✓	✓		✓				1	3	4	8
FourSquare	Liu et al. (2021)	Locations	✓							0	3	0	3
IJCAI2015	Tianchi (2018b)	Shopping		✓						0	3	0	3
Amazon Electronic	Wan et al. (2020)	Electronic Articles	✓							0	3	0	3
Sushi	Kamishima (2022)	Sushi	✓	✓						0	2	0	2
Speeddate	Fisman et al. (2006)	Dating Matches					✓			0	1	1	2
Facebook[c]	Kosinski et al. (2015)	College Majors	✓							0	2	0	2
Book-Crossing	Ziegler et al. (2005)	Books	✓	✓						0	2	0	2
Kiva[a]	Kiva (2022)	Micro-lending						✓[b]		0	1	0	1
Twitter, expert topic	Ge et al. (2016)	Experts + Topics					✓[b]			0	1	0	1
DBLP	Tang et al. (2008)	Co-authors						✓[b]		0	1	0	1
Insurance	Zindi (2022)	Insurance Products	✓		✓	✓				0	1	0	1
ModCloth	Wan et al. (2020)	Clothing							✓	0	1	0	1
MSN News	Wu et al. (2019)	News	✓							0	1	0	1
Instagram	Zhang et al. (2018)	Locations	✓	✓						0	1	0	1
MathNation[d]	MathNation (2022)	Learning Peers	✓				✓			0	1	0	1
CIKM 2019	Tianchi (2022)	E-commerce	✓	✓						0	1	0	1
Taobao Ad	Tianchi (2018a)	Ads	✓	✓						0	1	0	1

[a] The service offers an API that has been used to compile different datasets

[b] The sensitive attribute relates to the provider side, not the consumer side

[c] The dataset is no longer distributed

[d] The dataset is not publicly available

Table 12 Table representing the different datasets applied in fair consumer-side recommender systems research, without sensitive attributes

	Reference	Setting	Count			
			Pre-	In-	Post-	Total
Google Local Review	He et al. (2017a)	Locations	0	0	4	4
Flixster	Jamali and Ester (2010)	Films	1	2	0	3
Reddit[a]	Reddit (2022)	Forum Boards	0	1	1	2
Amazon	He and McAuley (2016)	E-commerce	0	1	1	2
Yelp Challenge	Yelp (2022)[b]	Locations	1	0	0	1
Freebase15k-237	Toutanova and Chen (2015)	Knowledge Base Completion	0	1	0	1
BeerAdvocate	McAuley et al. (2012)	Beers	0	1	0	1
DPG Recruitment[c]	DPG-Recruitment (2022)	Jobs	0	1	0	1
Twitter, scientific rumour	De Domenico et al. (2013)	Followers	0	0	1	1
Ctrip[c]	Ctrip (2022)	Flights	0	0	1	1

[a] The service offers an API that has been used to compile different datasets

[b] The variation of the dataset used in the Yelp challenge and by Fang et al. (2022), is no longer distributed

[c] The dataset is not publicly available

probabilistically model a “same religion" attribute for linking with an explicitly provided “same religion preference" attribute. Finally, Fang et al. (2022) derived gender labels of Yelp users Yelp (2022) based on their provided names.

9 Future directions

Given how new the field is, it is not easy to identify and recommend promising directions in terms of model architecture, etc. There is likely also not a single fairness definition that everyone will be able to agree on, so the field is undoubtedly going to continue exploring multiple parallel directions in the foreseeable future. However, regarding reproducibility aspects and other measures for improving the credibility of the research and approaches, various points could benefit from additional focus in the coming years. In particular, we perceive a need for consolidating fairness concepts, working towards standardizing the fairness metrics and improving comparisons with other approaches.

9.1 Consolidated fairness concepts and definitions

A recurring observation in the studies covered in this survey is the lack of a common language when it comes to fairness concepts and definitions. It often falls to the reader to interpret exactly what the authors consider fair by examining the text, the implementation choices and the evaluation. This survey highlights that there are multiple fairness definitions researched that differ significantly on a conceptual level and that are often conflicting in terms of optimization and goals. These factors complicate the examination of new research as well as comparisons of different models, and a common understanding of high-level fairness concepts could do much in to remedy such challenges. One may enhance the reader's ability to put new approaches, as well as implementation and evaluation choices, into context by immediately and accurately conveying the high-level fairness goal. In this case, the readers do not have to fully grasp the finer details and implications of the specific fairness definitions before they are able to make sense of the discussion and draw parallels with approaches they are familiar with. This may also assist researchers in identifying relevant research, and help structure further research while leaving room for more specific formal definitions that fall under the high-level fairness goals. The Fairness Interpretations taxonomy proposed in Sect. 3.2.1 is one suggestion for such high-level conceptual categories.

9.2 Consensus on Fairness metrics

Section 7 demonstrates a great number of applied fairness metrics and a high degree of overlap in what they essentially seek to measure. While this is natural for a budding field, and enhanced by the presence of multiple distinct and conflicting fairness definitions, it is currently a contributing factor in making the comparisons challenging. Guided by rigorous analysis of the properties of different metrics, the field as a whole could benefit from reducing the number of metrics applied by identifying the best among metrics that have higher degrees of overlap.

9.3 Comparison

Despite a growing number of studies covering similar fairness concepts, there is still a low degree of comparative analysis of different approaches. While it is interesting to see how fairness-aware contributions affect the fairness over the base recommender approaches, it is also essential to compare with relevant fairness-aware approaches, if present. This aspect seems to have improved recently, but there is still room for further improvement.

One contributing factor to the lack of comparative analysis is likely visibility. The research field is still relatively new, and the nomenclature has yet to consolidate, making it challenging to identify similar research. There is also an issue of visibility across different types of approaches, in particular recommender systems, IR Ranking and link-prediction. Both IR Ranking and link-prediction approaches may be considered recommender systems, depending on the setting or the applied dataset. However, since they use different terms than those used in the recommender system research and intermingling between fields can be uncommon, such approaches may not be known by researchers proposing similar recommender systems. Visibility has also been limited so far by the lack of updated surveys that chart out the field's current state. However, recent contributions like the comparative analysis in Boratto et al. (2022) and future surveys will hopefully improve this aspect.

9.4 Datasets

As noted in Sect. 8, there are currently not that many relevant datasets for evaluating consumer-side recommender systems fairness. A wider selection of benchmarking datasets could improve evaluation and comparisons and add credibility to the research. New datasets should ideally vary in size and sources to offer different challenges related to aspects like scalability and adaptability, focusing on filling in the gaps not covered by the datasets applied today. In particular, many current datasets are getting old, and their application may fail to reflect performance in a shifting environment. Finally, to better highlight the need and application of fair recommender systems, it would be useful to have datasets for which the ethical implications of a discriminatory recommender system are more severe.

9.5 Online evaluation

None of the considered studies performs an online evaluation of neither recommendation utility nor fairness. While offline evaluation has some practical benefits, it is usually restricted to only being able to reward recommendation of items/actions we know the user likes, not serendipitous recommendations of items the user will like but was not aware of when the dataset was created. Online A/B testing, on the other hand, can reward such recommendations and may bring along other benefits of testing the model in the environment it will be used, granted that they are designed and executed well. Further, online evaluation allows more subjective feedback, e.g., asking the users if they suspect that the recommender system discriminates against them or presents them with biased recommendations influenced by their inferred or stated sensitive attributes.

While research like Saxena et al. (2019) looks into the public pre-conceived perception and attitude towards formal fairness definitions, the impression of those using a fairness-aware recommender system may differ. Multiple approaches covered in this survey strive to make their models or recommendations independent of sensitive attributes. It would be interesting to see how different users perceive such a system in different recommendation settings.

Author contributions BV performed the selection process and wrote the main manuscript text. All authors read and discussed the selected studies, and all authors reviewed the manuscript.

Funding Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital). This publication has been partly funded by the SFI NorwAI, (Centre for Research-based Innovation, 309834). The authors gratefully acknowledge the financial support from the Research Council of Norway and the partners of the SFI NorwAI.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Afsar MM, Crump T, Far B (2022) Reinforcement learning based recommender systems: a survey. ACM Comput Surv 55(7):1–38. https://doi.org/10.1145/3543846

Ashokan A, Haas C (2021) Fairness metrics and bias mitigation strategies for rating predictions. Inf Process Manag 58(5):102646. https://doi.org/10.1016/j.ipm.2021.102646

Bach SH, Broecheler M, Huang B et al (2017) Hinge-loss markov random fields and probabilistic soft logic. J Mach Learn Res 18(1):3846–3912

Bhattacharyya A (1943) On a measure of divergence between two statistical populations defined by their probability distributions. Bull Calcutta Math Soc 35:99–109

Biswas A, Patro GK, Ganguly N et al (2021) Toward fair recommendation in two-sided platforms. ACM Trans Web 16(2):1–34. https://doi.org/10.1145/3503624

Bobadilla J, Lara-Cabrera R, González-Prieto A et al (2021) DeepFair: deep learning for improving fairness in recommender systems. Int J Interact Multimed Artif Intell 6(6):86. https://doi.org/10.9781/ijimai.2020.11.001

Boratto L, Fenu G, Marras M et al (2022) Consumer fairness in recommender systems: contextualizing definitions and mitigations. In: Hagen M, Verberne S, Macdonald C et al (eds) Advances in Information Retrieval. Springer International Publishing, Cham, pp 552–566

Borges R, Stefanidis K (2022) F2VAE: a framework for mitigating user unfairness in recommendation systems. In: Proceedings of the 37th ACM/SIGAPP symposium on applied computing. ACM, Virtual Event, pp 1391–1398. https://doi.org/10.1145/3477314.3507152

Bose A, Hamilton W (2019) Compositional fairness constraints for graph embeddings. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, proceedings of machine learning research, vol 97. PMLR, pp 715–724. URL https://proceedings.mlr.press/v97/bose19a.html

Breiman L (2001) Random forests. Mach Learn 45(1):5–32. https://doi.org/10.1023/A:1010933404324

Budish E (2011) The combinatorial assignment problem: approximate competitive equilibrium from equal incomes. J Polit Econ 119(6):1061–1103. https://doi.org/10.1086/664613

Burke R, Sonboli N, Ordonez-Gauger A (2018) Balanced neighborhoods for multi-sided fairness in recommendation. In: Friedler SA, Wilson C (eds) Proceedings of the 1st conference on fairness, accountability and transparency, proceedings of machine learning research, vol 81. PMLR, pp 202–214. URL https://proceedings.mlr.press/v81/burke18a.html

Buyl M, Bie TD (2020) DeBayes: a Bayesian method for debiasing network embeddings. In: Proceedings of the 37th international conference on machine learning, ICML 2020, 13–18 July 2020, Virtual Event, Proceedings of machine learning research, vol 119. PMLR, pp 1220–1229. URL http://proceedings.mlr.press/v119/buyl20a.html

Caton S, Haas C (2023) Fairness in machine learning: a survey. ACM Comput Surv. https://doi.org/10.1145/3616865

Ctrip (2022) Ctrip homepage. URL https://ctrip.com/, accessed: 2022-02-09

De Domenico M, Lima A, Mougel P et al (2013) The anatomy of a scientific rumor. Sci Rep 3(1):2980. https://doi.org/10.1038/srep02980

Deldjoo Y, Anelli VW, Zamani H et al (2021) A flexible framework for evaluating user and item fairness in recommender systems. User Model User-Adap Interact 31(3):457–511. https://doi.org/10.1007/s11257-020-09285-1

Deldjoo Y, Jannach D, Bellogin A et al (2023) Fairness in recommender systems: research landscape and future directions. User Model User-Adapt Interact. https://doi.org/10.1007/s11257-023-09364-z

Dickens C, Singh R, Getoor L (2020) HyperFair: a soft approach to integrating fairness criteria. In: 3rd FAccTRec workshop: responsible recommendation. p 10

Do V, Corbett-Davies S, Atif J et al (2021) Two-sided fairness in rankings via Lorenz dominance. In: Ranzato M, Beygelzimer A, Dauphin Y, et al (eds) Advances in neural information processing systems, vol 34. Curran Associates, Inc., pp 8596–8608. URL https://proceedings.neurips.cc/paper/2021/file/48259990138bc03361556fb3f94c5d45-Paper.pdf

DPG-Recruitment (2022) DPG recruitment homepage. URL https://www.dpgrecruitment.nl/, accessed: 2022-02-11

Dwork C (2011) Differential privacy, Springer US, Boston, pp 338–340. https://doi.org/10.1007/978-1-4419-5906-5_752

Edizel B, Bonchi F, Hajian S et al (2020) FaiRecSys: mitigating algorithmic bias in recommender systems. Int J Data Sci Anal 9(2):197–213. https://doi.org/10.1007/s41060-019-00181-5

Ekstrand MD, Das A, Burke R et al (2022) Fairness in recommender systems. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer US, New York, pp 679–707. https://doi.org/10.1007/978-1-0716-2197-4_18

Fang M, Liu J, Momma M et al (2022) FairRoad: achieving fairness for recommender systems with optimized antidote data. In: Dietrich S, Chowdhury O, Takabi D (eds) SACMAT '22: The 27th ACM symposium on access control models and technologies, New York, NY, USA, June 8–10, 2022. ACM, pp 173–184, https://doi.org/10.1145/3532105.3535023

Farnadi G, Kouki P, Thompson SK et al (2018) A fairness-aware hybrid recommender system. In: 2nd FATREC workshop. arXiv:1809.09030

Fisman R, Iyengar S, Kamenica E et al (2006) Gender differences in mate selection: evidence from a speed dating experiment. Q J Econ 121:673–697. https://doi.org/10.1162/qjec.2006.121.2.673

Frank M, Wolfe P (1956) An algorithm for quadratic programming. Naval Res Logist Q 3(1–2):95–110. https://doi.org/10.1002/nav.3800030109

Friedman A, Schuster A (2010) Data mining with differential privacy. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining. Association for computing machinery, New York, NY, USA, KDD '10, p 493–502. https://doi.org/10.1145/1835804.1835868

Frisch G, Leger JB, Grandvalet Y (2021) Co-clustering for Fair Recommendation. In: Kamp M, Koprinska I, Bibal A et al (eds) Machine Learning and Principles and Practice of Knowledge Discovery in Databases. Springer International Publishing, Cham, pp 607–630

Gajane P (2017) On formalizing fairness in prediction with machine learning. CoRR arXiv:abs/1710.03184

Gardner J, Brooks C, Baker R (2019) Evaluating the fairness of predictive student models through slicing analysis. In: Proceedings of the 9th international conference on learning analytics & knowledge. Association for computing machinery, New York, NY, USA, LAK19, p 225–234. https://doi.org/10.1145/3303772.3303791

Ge H, Caverlee J, Lu H (2016) Taper: a contextual tensor-based approach for personalized expert recommendation. In: Proceedings of the 10th ACM conference on recommender systems. Association for computing machinery, New York, NY, USA, RecSys '16, p 261–268. https://doi.org/10.1145/2959100.2959151

Govaert G, Nadif M (2008) Block clustering with Bernoulli mixture models: comparison of different approaches. Comput Stat Data Anal 52(6):3233–3245. https://doi.org/10.1016/j.csda.2007.09.007

Harper FM, Konstan JA (2015) The movielens datasets: history and context. ACM Trans Interact Intell Syst 5(4):1–19. https://doi.org/10.1145/2827872

Harrison G, Hanson J, Jacinto C et al (2020) An empirical study on the perceived fairness of realistic, imperfect machine learning models. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for computing machinery, New York, NY, USA, FAT* '20, p 392–402. https://doi.org/10.1145/3351095.3372831

He R, McAuley J (2016) Ups and downs: modeling the visual evolution of fashion trends with one-class collaborative filtering. In: Proceedings of the 25th international conference on world wide web. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, WWW '16, p 507–517. https://doi.org/10.1145/2872427.2883037

He R, Kang WC, McAuley J (2017a) Translation-based recommendation. In: Proceedings of the eleventh ACM conference on recommender systems. Association for computing machinery, New York, NY, USA, RecSys '17, p 161–169. https://doi.org/10.1145/3109859.3109882

He X, Liao L, Zhang H et al (2017b) Neural collaborative filtering. In: Proceedings of the 26th international conference on world wide web. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, WWW '17, p 173–182. https://doi.org/10.1145/3038912.3052569

Houlsby N, Giurgiu A, Jastrzebski S et al (2019) Parameter-efficient transfer learning for NLP. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, proceedings of machine learning research, vol 97. PMLR, pp 2790–2799. URL https://proceedings.mlr.press/v97/houlsby19a.html

Huang W, Labille K, Wu X et al (2021) Fairness-aware bandit-based recommendation. In: 2021 IEEE international conference on big data (big data). pp 1273–1278. https://doi.org/10.1109/BigData52589.2021.9671959

Islam R, Keya KN, Pan S et al (2019) Mitigating demographic biases in social media-based recommender systems. KDD (Social Impact Track)

Islam R, Keya KN, Zeng Z et al (2021) Debiasing career recommendations with neural fair collaborative filtering. In: Proceedings of the web conference 2021. ACM, Ljubljana Slovenia, pp 3779–3790. https://doi.org/10.1145/3442381.3449904

Jamali M, Ester M (2010) A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on recommender systems. Association for computing machinery, New York, NY, USA, RecSys '10, p 135–142. https://doi.org/10.1145/1864708.1864736

Jannach D, Zanker M, Felfernig A et al (2010) Recommender systems: an introduction. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511763113

Järvelin K, Kekäläinen J (2002) Cumulated gain-based evaluation of IR techniques. ACM Trans Inf Syst 20(4):422–446. https://doi.org/10.1145/582415.582418

Kamishima T (2022) Sushi dataset. URL https://www.kamishima.net/sushi/, accessed: 2022-02-09

Kamishima T, Akaho S (2017) Considerations on Recommendation Independence for a Find-Good-Items Task. In: Workshop on responsible recommendation. https://doi.org/10.18122/B2871W

Kamishima T, Akaho S, Asoh H et al (2012) Enhancement of the neutrality in recommendation. In: Gemmis Md, Felfernig A, Lops P et al (eds) Proceedings of the 2nd workshop on human decision making in recommender systems, Dublin, Ireland, September 9, 2012, CEUR workshop proceedings, vol 893. CEUR-WS.org, pp 8–14. URL http://ceur-ws.org/Vol-893/paper2.pdf

Kamishima T, Akaho S, Asoh H et al (2013) Efficiency improvement of neutrality-enhanced recommendation. In: Chen L, Gemmis Md, Felfernig A et al (eds) Proceedings of the 3rd workshop on human decision making in recommender systems in conjunction with the 7th ACM conference on recommender systems (RecSys 2013), Hong Kong, China, October 12, 2013, CEUR workshop proceedings, vol 1050. CEUR-WS.org, pp 1–8. URL http://ceur-ws.org/Vol-1050/paper1.pdf

Kamishima T, Akaho S, Asoh H et al (2016) Model-based approaches for independence-enhanced recommendation. In: 2016 IEEE 16th international conference on data mining workshops (ICDMW). IEEE, Barcelona, Spain, pp 860–867. https://doi.org/10.1109/ICDMW.2016.0127

Kamishima T, Akaho S, Asoh H et al (2018) Recommendation independence. In: Friedler SA, Wilson C (eds) Conference on fairness, accountability and transparency, FAT 2018, 23–24 February 2018, New York, NY, USA, Proceedings of machine learning research, vol 81. PMLR, pp 187–201. URL http://proceedings.mlr.press/v81/kamishima18a.html

Kang B, Lijffijt J, Bie TD (2019) Conditional network embeddings. In: 7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net. URL https://openreview.net/forum?id=ryepUj0qtX

Kingma DP, Welling M (2022) Auto-encoding variational bayes. arXiv:1312.6114

Kiva (2022) Kiva homepage. URL https://www.kiva.org/, accessed: 2022-02-09

Koren Y, Rendle S, Bell R (2022) Advances in collaborative filtering. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer US, New York, pp 91–142. https://doi.org/10.1007/978-1-0716-2197-4_3

Kosinski M, Matz S, Gosling S et al (2015) Facebook as a research tool for the social sciences. Am Psychol 70:543–556. https://doi.org/10.1037/a0039210

Kouki P, Fakhraei S, Foulds J et al (2015) Hyper: a flexible and extensible probabilistic framework for hybrid recommender systems. In: Proceedings of the 9th ACM conference on recommender systems. Association for computing machinery, New York, NY, USA, RecSys '15, p 99–106. https://doi.org/10.1145/2792838.2800175

Langseth H, Nielsen TD (2012) A latent model for collaborative filtering. Int J Approx Reason 53(4):447–466. https://doi.org/10.1016/j.ijar.2011.11.002

Last.fm (2022) Last.fm homepage. URL https://www.last.fm/, accessed: 2022-02-09

Li C, Xing W, Leite WL (2022) Toward building a fair peer recommender to support help-seeking in online learning. Distance Educ 43(1):30–55. https://doi.org/10.1080/01587919.2021.2020619

Li CT, Hsu C, Zhang Y (2022) FairSR: fairness-aware sequential recommendation through multi-task learning with preference graph embeddings. ACM Trans Intell Syst Technol 13(1):1–21. https://doi.org/10.1145/3495163

Li RZ, Urbano J, Hanjalic A (2021a) Leave no user behind: towards improving the utility of recommender systems for non-mainstream users. In: Proceedings of the 14th ACM international conference on web search and data mining. ACM, Virtual Event Israel, pp 103–111. https://doi.org/10.1145/3437963.3441769

Li S, Karatzoglou A, Gentile C (2016) Collaborative filtering bandits. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. Association for computing machinery, New York, NY, USA, SIGIR ’16, p 539–548. https://doi.org/10.1145/2911451.2911548

Li T, Sahu AK, Talwalkar A et al (2020) Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag 37(3):50–60. https://doi.org/10.1109/MSP.2020.2975749

Li X, She J (2017) Collaborative variational autoencoder for recommender systems. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. Association for computing machinery, New York, NY, USA, KDD ’17, p 305–314. https://doi.org/10.1145/3097983.3098077

Li Y, Chen H, Xu S et al (2021b) Towards personalized fairness based on causal notion. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. ACM, Virtual Event Canada, pp 1054–1063. https://doi.org/10.1145/3404835.3462966

Li Y, Chen H, Xu S et al (2023) Fairness in recommendation: foundations, methods, and applications. ACM Trans Intell Syst Technol 14(5):1–48. https://doi.org/10.1145/3610302

Liu H, Lin H, Fan W et al (2022) Self-supervised learning for fair recommender systems. Appl Soft Comput 125:109126. https://doi.org/10.1016/j.asoc.2022.109126

Liu H, Wang Y, Lin H et al (2022) Mitigating sensitive data exposure with adversarial learning for fairness recommendation systems. Neural Comput Appl 34(20):18097–18111. https://doi.org/10.1007/s00521-022-07373-4

Liu H, Zhao N, Zhang X et al (2022) Dual constraints and adversarial learning for fair recommenders. Knowl-Based Syst 239(108):058. https://doi.org/10.1016/j.knosys.2021.108058

Liu Q, Mu L, Sugumaran V et al (2021) Pair-wise ranking based preference learning for points-of-interest recommendation. Knowl-Based Syst 225(107):069. https://doi.org/10.1016/j.knosys.2021.107069

Masthoff J, Delić A (2022) Group recommender systems: beyond preference aggregation. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer US, New York, pp 381–420. https://doi.org/10.1007/978-1-0716-2197-4_10

MathNation (2022) Mathnation homepage. URL https://www.mathnation.com/, accessed: 2022-02-09

McAuley J, Leskovec J, Jurafsky D (2012) Learning attitudes and attributes from multi-aspect reviews. In: 2012 IEEE 12th international conference on data mining. pp 1020–1025

Mehrabi N, Morstatter F, Saxena N et al (2021) A survey on bias and fairness in machine learning. ACM Comput Surv 54(6):1–35. https://doi.org/10.1145/3457607

Narayanan A, Shmatikov V (2008) Robust de-anonymization of large sparse datasets. In: 2008 IEEE symposium on security and privacy (sp 2008), pp 111–125. https://doi.org/10.1109/SP.2008.33

Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Front Neurorobot 7:21. https://doi.org/10.3389/fnbot.2013.00021

Nikolakopoulos AN, Ning X, Desrosiers C et al (2022) Trust your neighbors: a comprehensive survey of neighborhood-based methods for recommender systems. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer US, New York, pp 39–89. https://doi.org/10.1007/978-1-0716-2197-4_2

Page MJ, McKenzie JE, Bossuyt PM et al (2021) The prisma 2020 statement: an updated guideline for reporting systematic reviews. BMJ. https://doi.org/10.1136/bmj.n71

Paraschakis D, Nilsson BJ (2020) Matchmaking under fairness constraints: a speed dating case study. In: Boratto L, Faralli S, Marras M et al (eds) Bias and social aspects in search and recommendation. Springer International Publishing, Cham, pp 43–57

Patro GK, Biswas A, Ganguly N et al (2020a) FairRec: two-sided fairness for personalized recommendations in two-sided platforms. In: Proceedings of world wide web conference (WWW 2020). ACM, pp 1194–1204. https://doi.org/10.1145/3366423.3380196

Patro GK, Chakraborty A, Ganguly N et al (2020) Incremental fairness in two-sided market platforms: on smoothly updating recommendations. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, no 01. pp 181–188. https://doi.org/10.1609/aaai.v34i01.5349

Pitoura E, Stefanidis K, Koutrika G (2022) Fairness in rankings and recommendations: an overview. VLDB J 31(3):431–458. https://doi.org/10.1007/s00778-021-00697-y

ProPublica (2016) Machine bias. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Rahman T, Surma B, Backes M et al (2019) Fairwalk: towards fair graph embedding. In: Proceedings of the twenty-eighth international joint conference on artificial intelligence. International joint conferences on artificial intelligence organization, Macao, China, pp 3289–3295. https://doi.org/10.24963/ijcai.2019/456

Rastegarpanah B, Gummadi KP, Crovella M (2019) Fighting fire with fire: using antidote data to improve polarization and fairness of recommender systems. In: Proceedings of the twelfth ACM international conference on web search and data mining. ACM, Melbourne VIC Australia, pp 231–239. https://doi.org/10.1145/3289600.3291002

Reddit (2022) Reddit homepage. URL https://www.reddit.com/, accessed: 2022-02-09

Resheff YS, Elazar Y, Shahar M et al (2019) Privacy and fairness in recommender systems via adversarial training of user representations. In: Marsico MD, Baja GSd, Fred ALN (eds) Proceedings of the 8th international conference on pattern recognition applications and methods, ICPRAM 2019, Prague, Czech Republic, February 19–21, 2019. SciTePress, pp 476–482. https://doi.org/10.5220/0007361204760482

Ricci F, Rokach L, Shapira B (2022) Recommender systems handbook. Springer US, Boston. https://doi.org/10.1007/978-1-0716-2197-4

Ricci F, Rokach L, Shapira B (2022) Recommender systems: techniques, applications, and challenges. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer US, New York, pp 1–35. https://doi.org/10.1007/978-1-0716-2197-4_1

Rus C, Luppes J, Oosterhuis H et al (2022) Closing the gender wage gap: adversarial fairness in job recommendation. In: RecSys in HR'22: the 2nd workshop on recommender systems for human resources, in conjunction with the 16th ACM conference on recommender systems. https://doi.org/10.48550/arXiv.2209.09592, arXiv:2209.09592

Saxena NA, Huang K, DeFilippis E et al (2019) How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. Association for computing machinery, New York, NY, USA, AIES '19, p 99–106. https://doi.org/10.1145/3306618.3314248

Slokom M, Hanjalic A, Larson M (2021) Towards user-oriented privacy for recommender system data: a personalization-based approach to gender obfuscation for user profiles. Inf Process Manag 58(6):102,722. https://doi.org/10.1016/j.ipm.2021.102722

Sonboli N, Smith JJ, Cabral Berenfus F et al (2021) Fairness and transparency in recommendation: the users' perspective. In: Proceedings of the 29th ACM conference on user modeling, adaptation and personalization. Association for computing machinery, New York, NY, USA, UMAP '21, p 274–279. https://doi.org/10.1145/3450613.3456835

Steck H (2018) Calibrated recommendations. In: Proceedings of the 12th ACM conference on recommender systems. Association for computing machinery, New York, NY, USA, RecSys '18, p 154–162. https://doi.org/10.1145/3240323.3240372

Tang J, Zhang J, Yao L et al (2008) Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. Association for computing machinery, New York, NY, USA, KDD '08, p 990–998. https://doi.org/10.1145/1401890.1402008

Tianchi (2018a) Ad display/click data on taobao.com. URL https://tianchi.aliyun.com/dataset/dataDetail?dataId=56, accessed: 2022-02-09

Tianchi (2018b) IJCAI-15 repeat buyers prediction dataset. URL https://tianchi.aliyun.com/dataset/dataDetail?dataId=42, accessed: 2022-02-09

Tianchi (2022) CIKM 2019. URL https://tianchi.aliyun.com/competition/entrance/231719/introduction/, accessed: 2022-10-10

Toutanova K, Chen D (2015) Observed versus latent features for knowledge base and text inference. In: Proceedings of the 3rd workshop on continuous vector space models and their compositionality. Association for computational linguistics, Beijing, China, pp 57–66. https://doi.org/10.18653/v1/W15-4007

Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, et al. (eds) Advances in neural information processing systems, vol 30. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wan M, Ni J, Misra R et al (2020) Addressing marketing bias in product recommendations. In: Proceedings of the 13th international conference on web search and data mining. Association for computing machinery, New York, NY, USA, WSDM '20, pp 618–626. https://doi.org/10.1145/3336191.3371855

Wang S, Hu L, Wang Y et al (2021) Graph learning based recommender systems: a review. In: Zhou ZH (ed) Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21. International joint conferences on artificial intelligence organization, pp 4644–4652. https://doi.org/10.24963/ijcai.2021/630

Wang Y, Ma W, Zhang M et al (2022) A survey on the fairness of recommender systems. ACM Trans Inf Syst. https://doi.org/10.1145/3547333

Wei T, He J (2022) Comprehensive fair meta-learned recommender system. In: Zhang A, Rangwala H (eds) KDD '22: the 28th ACM SIGKDD conference on knowledge discovery and data mining, Washington, DC, USA, August 14–18, 2022. ACM, pp 1989–1999. https://doi.org/10.1145/3534678.3539269

Weinsberg U, Bhagat S, Ioannidis S et al (2012) Blurme: inferring and obfuscating user gender based on ratings. In: Proceedings of the sixth ACM conference on recommender systems. Association for computing machinery, New York, NY, USA, RecSys '12, p 195–202. https://doi.org/10.1145/2365952.2365989

Wu C, Wu F, Qi T et al (2019) Neural gender prediction from news browsing data. In: Sun M, Huang X, Ji H et al (eds) Chinese Computational Linguistics. Springer International Publishing, Cham, pp 664–676

Wu C, Wu F, Wang X et al (2021) Fairness-aware news recommendation with decomposed adversarial learning. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, no 5. pp 4462–4469. https://doi.org/10.1609/aaai.v35i5.16573

Wu H, Ma C, Mitra B et al (2022) A multi-objective optimization framework for multi-stakeholder fairness-aware recommendation. ACM Trans Inf Syst. https://doi.org/10.1145/3564285

Wu L, Chen L, Shao P et al (2021b) Learning fair representations for recommendation: a graph-based perspective. In: Proceedings of the web conference 2021. Association for computing machinery, New York, NY, USA, WWW '21, pp 2198–2208. https://doi.org/10.1145/3442381.3450015

Wu Y, Cao J, Xu G et al (2021c) TFROM: a two-sided fairness-aware recommendation model for both customers and providers. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. Association for computing machinery, New York, NY, USA, p 1013–1022. URL https://doi.org/10.1145/3404835.3462882

Wu Y, Xie R, Zhu Y et al (2022b) Selective fairness in recommendation via prompts. In: Amigó E, Castells P, Gonzalo J, et al. (eds) SIGIR '22: the 45th international ACM SIGIR conference on research and development in information retrieval, Madrid, Spain, July 11–15, 2022. ACM, pp 2657–2662. https://doi.org/10.1145/3477495.3531913

Xu B, Cui Y, Sun Z et al (2021) Fair Representation Learning in knowledge-aware recommendation. In: 2021 IEEE international conference on big knowledge (ICBK), pp 385–392. https://doi.org/10.1109/ICKG52313.2021.00058

Yao S, Huang B (2017) Beyond parity: fairness objectives for collaborative filtering. In: Guyon I, Luxburg Uv, Bengio S, et al. (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp 2921–2930. URL https://proceedings.neurips.cc/paper/2017/hash/e6384711491713d29bc63fc5eeb5ba4f-Abstract.html

Yao S, Huang B (2021) Personalized regularization learning for fairer matrix factorization. In: Karlapalem K, Cheng H, Ramakrishnan N et al (eds) Advances in knowledge discovery and data mining. Springer International Publishing, Cham, pp 600–611

Yelp (2022) Yelp dataset. URL https://www.yelp.com/dataset/, accessed: 2022-10-10

Zhang S, Tay Y, Yao L et al (2022) Deep learning for recommender systems. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer US, New York, pp 173–210. https://doi.org/10.1007/978-1-0716-2197-4_5

Zhang Y, Humbert M, Rahman T et al (2018) Tagvisor: a privacy advisor for sharing hashtags. In: WWW '18: proceedings of the 2018 world wide web conference, pp 287–296. https://doi.org/10.1145/3178876.3186095

Zheng Y, Dave T, Mishra N et al (2018) Fairness in reciprocal recommendations: a speed-dating study. In: Mitrovic T, Zhang J, Chen L, et al. (eds) Adjunct publication of the 26th conference on user modeling, adaptation and personalization, UMAP 2018, Singapore, July 08–11, 2018. ACM, pp 29–34. https://doi.org/10.1145/3213586.3226207

Ziegler CN, McNee SM, Konstan JA et al (2005) Improving recommendation lists through topic diversification. In: Proceedings of the 14th international conference on world wide web. Association for computing machinery, New York, NY, USA, WWW '05, p 22–32. https://doi.org/10.1145/1060745.1060754

Zindi (2022) Zindi insurance dataset. URL https://www.kaggle.com/mrmorj/insurance-recommendation, accessed: 2022-02-09

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