Fairness in Recommender Systems

Michael D. Ekstrand, Anubrata Das, Robin Burke, and Fernando Diaz

Abstract Recent research has added the idea of fairness to the suite of concerns beyond accuracy or user satisfaction that recommender systems researchers and practitioners consider in their work. Recommender systems pose unique challenges for investigating the fairness and non-discrimination concepts that have been developed in other machine learning literature. The multistakeholder nature of recommender applications, the ranked outputs, the centrality of personalization, and the role of user response complicate the problem of identifying precisely what types and operationalizations of fairness may be relevant. In this chapter, we lay out the various ways a recommender system may be unfair and provide a conceptual framework for identifying the the fairness that arise in an application and designing a project to assess and mitigate them. We then survey of the literature to date on fair recommendation and provide pointers to other research on algorithmic fairness we believe is a promising basis for improving the fairness of recommender systems.

1 Introduction

Throughout the history of recommender systems, both research and development have examined a range of effects beyond the accuracy and user satisfaction of the

Michael D. Ekstrand
People and Information Research Team, Boise State University, e-mail: michaelekstrand@boisestate.edu

Anubrata Das
University of Texas at Austin, e-mail: anubrata@utexas.edu

Robin Burke
That Recommender Systems Lab, University of Colorado Boulder, e-mail: robin.burke@colorado.edu

Fernando Diaz
Microsoft Research, e-mail: diazf@acm.org

system. These effects include the *diversity* of recommendation lists (see Chapter DIVERSITY), *coverage* of the system (see Chapter EVALUATION), and *novelty* and/or *serendipity* of recommendations (see Chapter NOVELTY), among other concerns.

In recent years, this concern has extended to the *fairness* of the system: are the benefits and resources it provides fairly allocated between different people or organizations it affects? This challenge is connected to the broader set of research on fairness in sociotechnical systems generally and AI systems more particularly (Mitchell et al., 2020; Barocas et al., 2019), but information access systems, including recommender systems, have their own set of particular challenges and possibilities.

Fair recommendation is not an entirely new concern, but builds on concepts with long precedent in the recommender systems literature. Work on *popularity bias* (Cañamares and Castells, 2018), along with attempts to ensure quality and equity in *long-tail recommendations* (Ferraro, 2019), can be viewed as a fairness problem to prevent the system from inordinately favoring popular, well-known, and possibly well-funded content creators. Group recommendation (see Chapter GROUP REC) is also concerned with ensuring that the various members of a group are treated fairly (Kaya et al., 2020). The problems we discuss here go beyond these (important) problems to examine how various biases — particularly *social* biases — can make their way in to the data, algorithms, and outputs of recommender systems.

Our primary goal in this chapter is to guide the reader through pieces needed in order to measure and possibly improve the fairness of a recommender system. This starts with a clear picture of what, precisely, it means for the system to be unfair with respect to its stakeholders and problem space, and the harms this unfairness may cause. There are many different fairness problems in personalization and recommendation and each problem has its own particular features and solutions; therefore the details of measuring and providing fairness are as varied as the applications to which they apply. To provide guidance in this space, we discuss the key decisions and resources necessary to carry out a fair recommender systems experiment, grounded in a specific example from our own work, and then survey the survey the research to date in fair recommender systems to provide pointers for further reading.

Like much of recommender systems, fairness in recommendation has significant overlap with fairness in information retrieval and draws heavily from machine learning literature; researchers and practitioners looking to study or improve the fairness of recommendations will do well to pay attention to a broad set of research results.

1.1 Example Applications

Throughout this chapter, we will use the problems pace of job and candidate search to illustrate the challenges and possibilities for producing fair recommendations. Many online platforms attempt to connect job-seekers and employment opportunities in some way. Some of these are are dedicated employment-seeking platforms, while others, such as LinkedIn and Xing, are more general-purpose professional networking platforms for which job-seeking is one (important) component.

Job-seeking is a multi-sided problem — people need good employment, and employers need good candidates — and also has significant fairness requirements that are often subject to regulation in various jurisdictions. Some of the specific fairness concerns for this application include:

- Are recommendations of job opportunities distributed fairly across users?
- Are job/candidate fit scores fair, or do the under- or over-score certain candidates?
- Do users have a fair opportunity to appear in result lists when recruiters are looking for candidates for a job opening?
- Are employers in protected groups (minority-owned businesses, for example) having their jobs fairly promoted to qualified candidates?
- Are there specific fairness concerns that come from regulatory requirements?

Job search is certainly not the only application with fairness concerns, however. The recommendations in music platforms, such as Spotify and Pandora, connect listeners with artists and have significant impact both on the user's listening experience and on the artists' financial and career prospects. A system that does systematically under-promotes particular artists or groups of artists may reduce their visibility and with it harm the financial viability of their artistic endeavors.

1.2 Fundamental Concepts in (Un)Fairness

Algorithmic fairness in general is concerned with going beyond the aggregate accuracy or effectiveness of a system — often, but not always, a machine learning application — to studying the *distribution* of its positive or negative effects on its subjects. Much of this work has been focused on fairness in classification or scoring systems; Mitchell et al. (2020) catalog key concepts in fair classification, and Hutchinson and Mitchell (2019) situate these concepts in the broader history of fairness in educational testing where many similar ideas were previously developed.

There are many ways to break down the various concepts of algorithmic fairness that have been studied in the existing literature, which we summarize in Table 1. The first is *individual* versus *group* fairness. Individual fairness sets the constraint that *similar individuals should be treated similarly*: given a function that can measure the similarity of two individuals with respect to a task, such as the ability of job applicants to perform the duties of a job, individuals with comparable ability should receive comparable decisions (Dwork et al., 2012). This decision construct is often probabilistic, to allow for the ability to only select one applicant for any given job; individual fairness in such cases requires that the probability of acceptance be similar for similarly-qualified individuals. This definition depends on the availability of an unbiased comparison of individuals — a significant limitation — and places no requirements on the treatment of dissimilar subjects.

Group fairness considers the system's behavior with respect to the group membership or identities of subjects. Often, this is realized by dividing the observed features

Individual fairness	Similar individuals have similar experience
Group fairness	Different groups have similar experience
Sensitive attribute	Attribute identifying group membership
Disparate treatment	Groups explicitly treated differently
Disparate impact	Groups receive outcomes at different rates
Disparate mistreatment	Groups receive erroneous (adverse) effects at different rates
Distributional harm	Harm caused by (unfair) distribution of resources or outcomes.
Representational harm	Harm caused by inaccurate internal or external representation.
Anti-classification	Protected attribute should not play a role in decisions.
Anti-subordination	Decision process should actively work to undo past harm.

Table 1 Summary of concepts in algorithmic fairness and harms.

into two sets $\langle X_i, A_i \rangle$, where A_i is a set of *sensitive attributes* identifying group membership or identity, often something like race or ethnicity, and X_i is the other features or covariates. Such work often designates a *protected group*, designated by a particular value of A_i, that should be protected from adverse discriminatory decisions. Many of the concepts are often grounded in U.S. legal notions of anti-discrimination (Barocas and Selbst, 2016), but this grounding is not without limitations, such as the resulting technical framings losing connection to the legal concepts that motivated them (Xiang and Raji, 2019).

The goal of group fairness is to address the impact of group membership on final decision outcomes, but the precise formulations of this goal further divide into several distinct sub-concepts of discriminatory behaviors the system should avoid. *Disparate treatment* is when the sensitive attribute is a direct feature in the decision-making process. *Disparate impact* arises when the outcomes, such as loan approvals, are disparately distributed between groups (e.g. one racial group has a higher loan approval rate than another). It is measured with constructs such as demographic parity (e.g. a resume screening process accepts resumes at the same rate). Both disparate treatment and impact are directly inspired by U.S. anti-discrimination law; disparate impact is not on its own enough to conclusively prove discrimination, but it triggers enhanced scrutiny of a process Barocas and Selbst (2016). *Disparate mistreatment*, on the other hand, looks at the distribution not of decision outcomes but of errors: it arises, for example, when one group has a higher false negative rate than another (for example, when members of one racial group is more likely to be denied loans for which they are qualified). Another disparate mistreatment construct is one sometimes called *equality of opportunity* (Hardt et al., 2016): the constraint that positive classification be conditionally independent of protected group given the true label — that is, a qualified job candidate from one group is no more likely to be erroneously screened out than a candidate from another.

Orthogonally to group and individual fairness, Crawford (2017) divided harms connected to unfairness into two categories: *distributional* harms, where a resource or opportunity is inequitably distributed (as with most of the harms discussed so far), and *representational* harms, where people in the system are represented in a way that is inaccurate and systematically unfair.

The motivations of fairness constructs can also be categorized, and different motivations yield different assessments of their effectiveness and adequacy. U.S. anti-discrimination law is rooted in competing ideas of *anti-classification*, the idea that the influence of a protected attribute should not play a role in decisions, and *anti-subordination*, the idea that current decision-making processes should actively work to reverse the effects of historical patterns of discrimination (Barocas and Selbst, 2016). These motivations are often present, and sometimes explicitly invoked, in algorithmic fairness literature, although their proper application is not entirely clear (Xiang and Raji, 2019).

While it is common to talk about "fairness" as the goal of a system designed to improve equity, it is impossible to make a system that is demonstrably and universally "fair". Competing ideals of fairness, the needs of multiple stakeholders, and the fundamentally social and contestable nature of fairness (Selbst et al., 2019) mean that it is not a fully solvable problem or even a well-specified one.

More detailed treatment of algorithmic fairness in general is beyond the scope of this chapter. We refer the reader to Mitchell et al. (2020) and Barocas et al. (2019) for further (essential) reading.

1.3 Multisided Analysis of Recommendation

One significant way in which fair recommendation differs from more general fair machine learning is in the *multisided* nature of recommendation. In many machine learning problems, it is clear who should be treated fairly: the system is making decisions about subjects, such as job or loan applicants. In recommender systems, however, different groups of stakeholders stand to benefit (or lose) from the system's behavior, and assessing the fairness of a recommender system requires identifying the stakeholders who may be treated unfairly. Chapter MULTISIDED of this book provides an introduction to multistakeholder recommendation broadly. For the purposes of this chapter, it is sufficient to note that different stakeholders have different fairness concerns (Burke, 2017). These include:

- **Consumer fairness**, concerned with treating different users of the system fairly by ensuring that users receive comparable service and no group of users is systematically disadvantaged by the system. For example, a system might be more inaccurate for some groups of users rather than others.
- **Provider fairness**, concerned with treating the creators or providers of items fairly, for example by ensuring that musicians have equitable opportunity for their music to be recommended to users in a streaming music platform.
- **Subject fairness**, concerned with fair treatment of the people or entities that items are about. One example of a domain where this concern may arise is news recommendation, where recommendations that systematically under-represent topics or segments of society can be considered subject-unfair.

These stakeholders' fairness concerns are often considered one at a time, but some work seeks to jointly analyze or provide fairness for multiple stakeholders simultaneously. In this chapter, we organize our literature survey around the stakeholders each work primarily considers.

1.4 Unfairness in Recommendation

As noted in Section 1.2, it is impossible to make a system fair according to any universal definition. What *can* be done is to measure and avoid specific types of well-defined *un*fairness. Different stakeholder groups can be harmed by the system's unfairness in different ways; aligning with Section 1.2, this unfairness can be group or individual, and can result in harms of both distribution and representation.

If a system gives some of its users lower-quality recommendations than others (Ekstrand et al., 2018a; Mehrotra et al., 2017), particularly if it systematically gives one socially-salient group of users less useful results than another, then it exhibits unfairness towards consumers that results in a distributional harm (the resource of "relevant information" is not fairly distributed). It may cause representational harm to its consumers if it misrepresents their personal information or interests. It may be unfair towards its consumers if it provides comparably good recommendations, as measured by click-through rate or another suitable metric, but it gives qualitatively different results to different groups, such as being less likely to show women ads for higher-paying jobs (Ali et al., 2019).

While provider fairness concerns prevent a relatively straightforward set of harms, there is a lot of nuance in how, precisely, to define the concept, but a system harms producers in an unfair way if it is more likely to recommend content from some creators than others. One well-known example of this is *popularity bias* (Cañamares and Castells, 2018), where the system is more likely to recommend already-popular content so less-popular creators are not as able to make their content visible and obtain recognition and commercial return for their efforts.

But visibility can also be misallocated along other lines, such as systematically under-recommending content by creators of particular genders or ethnic groups. One potential cause of this discrepancy can be exogenous inequities in opportunity to create content, publicity efforts from publishers and labels, and other systemic societal discrimination; the recommender system reifies and reproduces these inequities as it learns from that data in pursuit of its objective function. Another potential cause is objective functions that do not align with broader social interest — for example, optimizing information distribution in a network to maximize the number of nodes reached may leave significant gaps in access to information (Fish et al., 2019), a particularly relevant concern for things such as public health awareness campaigns. Misallocation may also arise from feedback loops, where the recommender system learns from user data and amplifies small differences in potential utility into large differences in recommendation exposure.

Fairness Questions:

1. What is the recommendation problem under study?
2. What is the *social goal* of fairness in this setting?
   - What stakeholders are considered?
   - What harm is to be prevented?
3. How is that goal *operationalized* with specific data, methods, and metrics?
4. What are the findings?
5. How do these findings map back to the social goal?
6. What are the scope and limitations of these methods and results?

Table 2 Key decisions in a fair recommender systems project.

In Sections 3–6, we will survey work on each of these (and other) harms. One of the crucial elements of pursuing fairness in research or system development, however, is to be specific about the harms under consideration. Identifying the specific stakeholders in play, and the particular harm or unfairness that we wish to measure or avoid, allows research to make progress and contribute techniques that can be a part of building recommender systems that promote equity instead of reproducing discrimination.

2 Studying Fairness

As discussed in Sections 1.3 and 1.4, there is no one problem of recommender system fairness: there are different stakeholders towards which the system may be unfair, and different ways in which it may be unfair to them. Our survey of the literature in Sections 3–7 discusses examples of work on many different problems in and approaches to fair recommendation; in this section, we describe the overall structure of a fair recommendation project and the key decisions that need to be made, regardless of the specific problem the researcher or developer seeks to address; Table 2 summarizes these.

As a running example in this section, we use the study of the fairness of collaborative filters with respect to books' authors' gender identities by Ekstrand and Kluver (2021), describing the decisions made in that work for each component of the process. We do not claim that this is a perfect example, but it provides a useful context to discuss how these general decision points need to be resolved in the context of a specific study. Also, while we discourage a needless proliferation of new fairness concepts, each individual fair recommendation project will be different due to the intrinsic differences between stakeholders, harms, applications, and data affordances that shape a particular effort.

2.1 Defining Fairness

The first step in any inquiry is to define precisely the aspects of fairness under study. There are several elements to consider, including:

- The **stakeholders** who should be treated fairly
- The **harms** to quantify and/or reduce
- The **metrics** for measuring and/or optimizing away the harm

A project may target more than one type of unfairness, but each fairness construct needs to be clearly described and account for these dimensions. It is important to be precise about both the social goal of a fairness construct and the way that goal is operationalized into specific methods and metrics. For example:

- To measure and reduce the potential harm that musicians of one ethnicity are more likely to be recommended than musicians of another, we may enforce a constraint that "the probability a song appears in a playlist is independent of musician ethnicity" (Kamishima et al., 2018). We may also measure the exposure (Diaz et al., 2020) or attention (Biega et al., 2018) providers of different ethnicities receive, and look for absolute disparities (disparate impact) or disproportional disparities based on relevance (disparate mistreatment).
- To measure and reduce the potential representational harm that users are stereotyped by gender in the system's latent feature space, we may measure the accuracy of classifying a user's gender from their embedding and use adversarial learning to minimize this predictability while providing accurate recommendations (Beutel et al., 2017).

The common theme of these operationalizations is that we have taken a concern — reduce a particular harm — and translated it into a specific framework and metric or objective function. This process is inherently reductive, as the resulting metric does not capture everything about the concern. Decisions in the process prioritize some aspects of fairness over others; for example, seeking to make artist groups' exposure proportional to their combined relevance will allow for significant disparities in exposure if they are supported by disparities in relevance judgements (such as user response or ratings), even if the relevance judgements are unfair (due to biases in user preference, presentation biases that affect likelihood to click, or other factors that bias the collection of relevance data).

Such limitations and tradeoffs do not necessarily invalidate a specific fairness construct; indeed, no construct is without them, and there is no well-defined universal concept of fairness. Rather, they are limitations that need to be documented. Clearly describing the concerns, the metrics, their relationship, and the limitations that arise from that choice of metrics will enable readers of a particular research paper to better understand its findings, and apply and extend them in additional situations.

The author gender project focuses on provider fairness as experienced by consumers, and seeks to measure the harm that authors of a particular gender may be under-represented in the recommendation lists the system provides to its users, even

if users read may books by authors of that gender. It operationalizes this by measuring the "% Female" for a set L of books (either the books in a user's reading history or the books recommended to them by a collaborative filter):

$$\text{SetBias}(L) = \frac{\text{\# of female-authored books in } L}{\text{\# of known-gender books in } L}$$

This definition captures one particular form of representation, but does not do anything to assess representation along other dimensions (e.g. race or ethnicity), or even capture the full subtlety of gender representation (such as accounting for non-binary gender identities).

Mismatches between a metric and its ultimate goals are common, however, particularly while the field is young and we are still determining how to study fairness and the relative merits of different approaches. We advocate for incrementally improving on the practice of fair recommendation research; perfect alignment and operationalization are likely only achievable in the limit.[1]

2.2 Collecting Data

Another major challenge is to find appropriate data for studying the fairness objective under consideration. While data sets for studying recommendation abound, fairness research — particularly group fairness — often requires additional data that is more difficult to come by. Olteanu et al. (2019) provide a thorough treatment of issues in data collection and augmentation, but here we highlight some specific considerations for fair recommendation research.

Commonly-used data sets are often directly amenable to individual fairness research, because all that is needed is users and items, and possibly item authorship. No further data augmentation is typically necessary.

Group fairness, both for providers and consumers, requires group affiliations or sensitive attributes for the relevant stakeholder entities. Depending on the particular type of group in question, this data may also be sensitive and subject to legal restrictions or ethical concerns for its use and/or distribution. This is particularly true for personal attributes such as gender, sexual orientation, or ethnicity.

Some data sets provide attribute labels information directly. The MovieLens 1M data set (Harper and Konstan, 2015) includes self-described binary gender identities and age brackets for MovieLens users (subsequent data sets do not, as MovieLens stopped collecting this information from its users). The Last.FM 360K data set

[1] In the conference version of the author gender paper, Ekstrand et al. (2018b) described the goal of promoting equality of opportunity for book authors, but measured fairness through the proportional composition of ranked lists. Exposure-oriented metrics (Diaz et al., 2020; Biega et al., 2018) would be a more coherent way of advancing the stated goal. The journal version described here more clearly contextualizes the capabilities and implications of the methods employed, because we're continually advancing our own understanding of these practices as well.

collected by Celma (2010)[2] and the later LastFM 1B data set (Schedl, 2016) also include user gender and age.

In other cases, a group can be synthesized from available data. For example, Kamishima et al. (2018) looked at fairness with respect to the age of a movie in order to test algorithms intended to be useful for a broader set of fairness problems. This can also be done with genres, keywords, and other information included in a data set. Care is needed, however, because as Selbst et al. (2019) argue, we cannot assume techniques for ensuring fairness with respect to one attribute will automatically apply to other attributes even within the same stakeholder class (e.g. a technique that addresses gender discrimination may or may not be effective for racial discrimination and vice versa, depending on the way that discrimination has affected the data generating process), but it is a useful technique for testing out technical possibilities for future repurposing, subject to re-validation on the target problem.

Sometimes, multiple public data sets can be integrated in order to study a fairness problem. For many media domains, there is publicly-available data on content creators such as book authors and film actors; sometimes, this data is in machine-readable form amenable to integration. In the author gender project, Ekstrand and Kluver (2021) integrated book ratings from several sources with OpenLibrary, Library of Congress bibliographic records, and the Virtual Internet Authority File to obtain the gender identities of many book authors; the integration tools are publicly available.[3] Because linked author records with gender identities are not available for all books, this project had to limit the final analysis to books where the first author's gender could be uniquely identified; the "% Female" metric is computed over the subset of rated or recommended books for which the author gender identity is known. Further, the available data only records binary gender identities, limiting the study's ability to properly account for authors with transgender, non-binary, and other gender identities.

Data can also be augmented through crowdsourcing or professional annotation. The TREC Fair Ranking Track (Biega et al., 2020) uses NIST's professional assessors to obtain information about the authors of scholarly papers from publicly-availablie sources. This data source is expensive, but depending on the task and source of labor it can result in high-quality annotations. Annotation does require careful attention to defining the legitimate bases for assigning a label; as one example, the Program for Cooperative Cataloging (Billey et al., 2016) has developed recommended best practices for recording authors' gender identities.

Some authors have used inference techniques to impute demographic labels to users or content producers. Common ways of doing this include statistical gender recognition based on names or computer vision techniques for gender recognition based on portraits or profile pictures. We generally discourage using this source of data, as it is error-prone, subject to systemic biases (Buolamwini and Gebru, 2018),

[2] `https://www.upf.edu/web/mtg/lastfm360k`

[3] `https://bookdata.piret.info`

reductionistic (Hamidi et al., 2018), and fundamentally denies subjects control over their identities.[4]

Another approach sometimes employed is to generate synthetic data, simulating the relevant information needed to assess fairness. This is motivated by the challenges of acquiring data with the appropriate demographic and content characteristics for academic researchers. If it is possible to establish correlates between aspects of user's profiles and demographic characteristics, then demographic labels can be probabilistically associated with user profiles, providing labels for protected and unprotected groups. This approach was demonstrated using job recommendation data in Burke et al. (2018a).

Finally, recent work has started to look at evaluating fairness without access to linked demographic records, for example by using background data on label distributions to calibrate estimation from proxy attributes (Kallus et al., 2019). This can be successful in some cases; Kallus et al. provide methods to identify some of those cases, but the resulting estimates can have very broad error bounds. Hashimoto et al. (2018) provide algorithms for fair learning without access to demographic labels, but the minority groups protected may differ from the groups needing protection in a particular application; the implications of this discrepancy will be application- and domain-specific.

No matter the technique used, it is crucial to carefully document the data sources, integration or augmentation decisions, and known limitations of the data. Gebru et al. (2018) identify many questions that should be answered for any data set, but are particularly germane to the kinds of data needed for fairness research. Due to the length limitations of many publication venues, this may necessitate appendices or online supplements that document the data sources and strategies more thoroughly than a conference paper allows, but we recommend that the paper itself include sufficient detail to understand and assess the limitations.

2.3 Structuring Experiments

The structure of a recommendation fairness study itself tends to follow structures typical for other recommender evaluations (see Chapter EVAL), particularly studies focusing on non-accuracy analyses such as diversity (Chapter DIVERSITY). The recommender algorithms are trained on available data, produce recommendations for users, and the resulting recommendations are measured for their fairness (and often other properties). This can be done with both offline and online experimental protocols. The only change to existing experimental practice typically required to study fairness is the additional data and metrics.

The author gender study uses a typical offline experimental protocol with user-based sampling. One sample of users are used for tuning algorithm hyperparameters

[4] We have deliberately omitted citations to papers violating the guidance in this paragraph because our purpose is to critique a general trend, not to attack any specific paper. We cite work using these techniques elsewhere in this chapter where it makes a relevant contribution.

with MRR in a leave-one-out protocol, and another sample is used for the fairness study. For the experiment's main results, the authors measure the gender composition of users' reading and rating histories, recommendation lists produced for these users by several collaborative filters, and the linear relationship between input and output composition (measuring the extent to which the rankings produced by each algorithm reflect the user's historical tendency to read authors of one gender or another).

2.4 Reporting Results

The results of a fairness study need to be carefully described and clearly contextualized in light of the study's specific aims and the limitations of the data and metrics. Special care needs to be taken to understand the implications of data and experimental limitations on the results and their generalizability. It is also important not to over-claim generality; *portability* is one of the abstraction traps identified by Selbst et al. (2019, §2.2), as results and solutions on one fairness concern do not necessarily apply to another, even if they can both be represented with the same statistical abstraction.

As a practical matter, we recommend clearly separating the social goal and the specific operationalization in both the motivation and discussion of the results in a paper. Being explicit about how the fairness goal or construct is translated into data and measurements, and how specific quantitative results illuminate the social goal, helps reduce the risks of the formalism trap (Selbst et al., 2019, §2.3) and provides context for discussing limitations.

The author-gender project is a purely observational, correlational study, and reports findings in light of the capabilities of such methods. It also describes in detail the data set; the coverage and distribution of relevant labels in the data; and limitations of the data, fairness construct and operationalization, and experimental design.

3 Consumer Fairness

Consumer fairness is the aspect of a recommender system concerned with how recommendations impact consumers and sub-groups of consumers. For example, a marketplace recommending job postings to job seekers may wish to ensure that all its users receive good-quality recommendations and comparable access to job opportunities, particularly marginalized groups such as women and people from minority ethnic backgrounds; such objectives may even be required by law.

3.1 Individual fairness

Since recommender systems' results are personalized to each consumer, users are expected to get different results, we might expect that what counts as "similar" individuals for the purposes of individual fairness in recommendation might be quite fine-grained. For example, a job recommender might be taking into account the user's past history of clicks or views on the site as well as more typical job history or personal data commonly associated with a job application. Recommender systems, especially collaborative ones, are typically built on the assumption that individuals will similar user profiles should indeed receive similar recommendations, so we might expect that individual fairness is a strength of these systems.

However, user behavior is not typically what is understood in the discussion of similar individuals in the context of individual fairness. A person's profile as gathered through interaction with a recommender system might include information that is irrelevant to the match between user and item. To continue with the employment example, a female user might be less likely to click on job descriptions with male-oriented language (Hentschel et al., 2014), and this fact may be completely irrelevant to her qualifications. Still, it may be quite difficult to ensure individual fairness in a recommendation setting, as a complete profile of all relevant information that would be needed to compare individuals could be quite challenging to obtain. Almost by definition, if a recommender system, especially a collaborative one, is being deployed, it is because the application is one where user profiles — and similarity — is better obtained through an implicit representation than explicit features.

Therefore, while individual fairness is a worthy objective, we expect that developers trying to ensure consumer-side fairness in recommendation will be primarily interested in framing their fairness concerns relative to group fairness.

3.2 Group fairness

A general approach to group fairness is to consider the benefit or utility that a recommender system delivers to different groups of users. To operationalize this concern, we need to determine how to measure utility and how to compare between groups. In some recommendation domains, it may be important to consider objective qualities of an item in understanding its utility, such as the salary in a job listing. If protected group users receive on the whole lower salary listings, this could be considered unfair regardless of other personalization considerations. Thus, there may be some inherent utility associated with each item to be taken account in the overall utility computation.

Most approaches, however, use the utility construct used for measuring the system's effectiveness, such as an accuracy metric based on ratings, click history, or similar profile data. Some group-fairness evaluations use the predicted rating or click probability estimated by the recommender system itself for a more complete but ap-

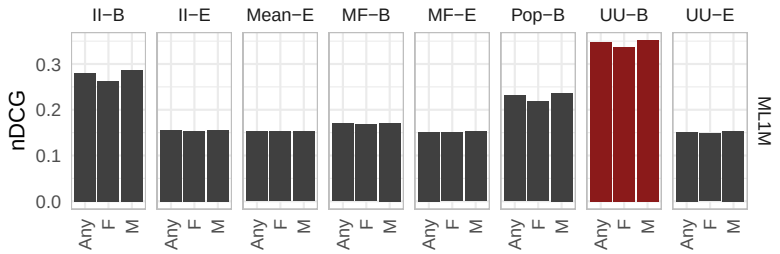

Fig. 1 Example results from a consumer group-fairness study by Ekstrand et al. (2018a). NDCG of algorithms trained on the MovieLens 1M data set is aggregated by user gender.

proximate picture of utility. With a known preference data, utility can be calculated in a number of ways (see also Chapter EVALUATION):

- Predicted rating fidelity: How closely do predicted ratings match known ones? The usual approach is to use RMSE or MAE to measure this accuracy.
- Predicted ranking accuracy: How closely do output rankings match known preferences? NDCG is a common utility-based measure for ranking.

Generally, experiments of this kind will look for disparate mistreatment, in the form of greater error or lower ranking quality for the protected group. One example that explores consumer-side fairness is by Ekstrand et al. (2018a). They performed an offline top-N evaluation of several collaborative filtering algorithms, computing NDCG and aggregating the results by user demographic. They found significant differences in utility between gender and (in some cases) user age groups, although not always in the same direction. Yao and Huang (2017) carried out similar work for prediction accuracy, developing several unfairness metrics over rating prediction errors for protected and unprotected groups. Typically-considered demographic groups are not the only basis for consumer fairness either; Abdollahpouri (2020) considers user groups based on their level of interest in popular items and shows that, across recommendation algorithms, users with an interest in less-popular niche items were not receiving recommendations in line with their interests.

3.3 Fairness Beyond Accuracy

While much of the work on consumer fairness focuses on the quality of recommendations, some work looks at other aspects of consumer experience that may be discriminatory, such as stereotyping or the specific items users receive. Ali et al. (2019) studied the distribution of ads on Facebook to understand potentially discriminatory impact in the visibility of different kinds of ads. They found that even when an advertiser wishes to have fair distribution of their ad, for example to ensure that an ad for a job opening is seen by people of all genders, the combination of

relevance optimization and market dynamics results in disparate distribution of ads across racial and gender lines. Nasr and Tschantz (2020) describe bidding strategies to attempt mitigate such effects and ensure fair ad distribution even when the platform does not provide it.

Beutel et al. (2017) present an approach to learning fair representations in a way that can be applied to consumers, by learning embeddings (such as the user and item embeddings in a recommender system) in an adversarial setting set up to minimize the ability to predict a user's sensitive attribute, such as gender, from their embedding. This has the potential to reduce stereotype effects in resulting recommendations.

3.4 More Complex Scenarios

The work described in this section makes assumptions that are often violated in realistic applications. One in particular is the assumption that only a single protected group or sensitive attribute needs to be considered for fairness. But our job recommender, for example, may need to meet constraints having to do with race, gender, religion, and other types of protected categories, as determined by applicable laws and organizational requirements. Legally, the complexities of the interaction of multiple protected categories has been explicated by Crenshaw (1989) and others under the framework of *intersectionality*. In the fair machine learning literature, it has been studied under the topic of *rich subgroup fairness* (Kearns et al., 2019). In recommender systems, there is some research involving sub-group fairness across providers (Sonboli et al., 2020b). However, no existing work addresses the compound nature of the disadvantage that Crenshaw highlights as characteristic for individuals who find themselves at the intersection of multiple protected identities.

Another simplification common in consumer-side fairness is assuming that categories are binary (protected vs unprotected) rather than constellations of attributes, including continuous qualities. Some work can naturally deal with multiple attributes — aggregating utility or error by group works fine for multiple groups, and multi-group statistical tests are effective — but optimization approaches typically assume that protected and unprotected groups should have equal results, and generalizing beyond two groups is seldom done.

4 Provider Fairness

Providers, the other side of the most common multistakeholder setup of recommender systems, are primarily the suppliers of information or items being recommended. There may be multiple parties who could be considered providers in most scenarios. For example, in news recommendation application, the journalist or publication venue (in a federated system) could be considered the provider. In music recommendation, it could be the individual artist or the record label. In job recommendation,

it could be the employer (when candidates are seeking jobs) or the prospective employee (when recruiters are looking for candidates).

The key benefit of a recommender system for providers is that the system provides exposure of their items to recommendation consumers who may be interested in those items. We can think of recommendation opportunities as a resource and the distribution of those opportunities across different groups as the characteristic provider fairness concern. For example, in the candidate recruitment recommendation scenario, when an employer is looking for candidates, different protected groups (e.g., gender and ethnic groups) should be treated fairly. Yet, the challenge remains to provide balanced outcomes between relevant and fairly distributed set of candidates.

Diversity (see Chapter DIVERSITY) is related to provider fairness. Some methods of diversity enhancement in recommendation may promote provider fairness, but diversity and fairness respond to different normative concerns. Diversity in recommendation and search results is mainly focused on consumer intent, by presenting results that meet a wide range of users' topical needs. In contrast, provider fairness is motivated by justice concerns to ensure that different providers receive fair opportunity for their content or products to be discovered.

4.1 Provider Utility

In contrast to consumer-side fairness, which is concerned with the quality of recommendations delivered to end-users, for provider-side fairness we are concerned with the utility of recommendations to providers. As noted above, a key notion in provider-side fairness is *exposure*: the value of recommendation opportunities given to a particular item or group of items.

Because recommendation lists are often short, a single list contains only a small number of recommendation opportunities; further, due to user attention patterns, not all list positions have equal utility, and only one item can have the benefits of the most valuable ranking in any one ranking (Diaz et al., 2020). Therefore, provider utility must usually be measured cumulatively (or in some cases amortized) over a sequence of recommendation results delivered to users. This is in contrast to the consumer-side case, where we can often meaningfully evaluate the quality of results for each user by looking only at a single recommendation result.

There are a number of options for computing the exposure / utility associated with a given recommendation opportunity. It may be convenient to assign a fixed utility to all items in a given recommendation list. In this case, the utility accrued to an item is simply the hit rate of its appearance across the recommendation experiment (Liu et al., 2019). Utility can also be weighted with a function estimating user attention to an item, which will vary greatly based on its ranking in a recommendation list. This is commonly represented through a logarithmic discount for higher rank items, as captured in the standard NDCG metric, or a geometric discount (Biega et al., 2018); for example, the exposure of an item i in a collection $\mathcal{L}$ of rankings under a geometric decay model can be computed by:

$$\text{Ex}_{\text{geom}}(i|\mathcal{L}) = \sum_{L \in \mathcal{L}} \gamma^{\text{rank}_L(i)}$$

4.2 Individual Fairness

The individual fairness principle for producer fairness says that "similar" items (or content producers) should receive similar utility from the recommender system. Defining similarity with respect to the recommendation task, however, requires some subtlety, particularly since each user will like a different set of items (as opposed to nonpersonalized classification settings, where items are similar with respect to a single task such as repaying a loan). One straightforward approach is to define "similar" based on relevance to the user: two items are similar if the same user likes them both. Under this paradigm, we can compare the exposure an item receives to the relevance or utility it provides to users; Biega et al. (2018) do this by aggregating each item's attention and relevance over multiple rankings, and measuring the extent to which attention is proportional to relevance (user-centric utility). Diaz et al. (2020) use relevance to determine an *ideal exposure* — the exposure items would receive if the system ranked the items perfectly with respect to user utility — and computes the difference between that exposure and the exposure under the actual system:

$$\text{EEL} = \sum_i \left(\text{Ex}_{\text{geom}}(i|\mathcal{L}) - \text{Ex}_{\text{geom}}(i|\mathcal{L}_{\text{ideal}})\right)^2$$

Another way to think about individual fairness is through latent representation. If individual items in the latent space are clustered together, then the downstream task of recommendation would preserve such characteristics (Lahoti et al., 2019). However, there is open work to explore how to formulate and achieve such representation specifically in the context of recommender systems by taking the trade-offs between personalization and fairness into consideration.

4.3 Group Fairness

As with consumer fairness and general ML fairness, provider group fairness seeks to ensure that content providers in different groups are treated fairly. In recommender systems, this needs to be connected to system's personalization goal: recommending items a user does not like to achieve a group-fairness goal is usually not appropriate.

One way to operationalize group-fairness is to measure or enforce statistical independence between the recommendation outcomes and protected attributes (Kamishima et al., 2018). Under this definition, introduced in Section 2.1, the recommendations are fair if the probability an item will appear in a particular recommendation list is independent of its sensitive attribute(s).

Another is to measure the extent to which protected groups are equitably represented in recommendation lists. This can be done by looking at the composition of individual recommendation rankings, measuring the divergence between the distribution of groups within a list and a target distribution (Yang and Stoyanovich, 2017; Das and Lease, 2019; Burke et al., 2018b). Given a recommendation list L, these approaches compute a distribution over protected groups $P(A|L)$ and a target $P_\tau(A)$; Sapiezynski et al. (2019) call this target a *population estimator*. Given these two distributions, they can be compared using difference in probabilities, odds ratios, KL divergence (Yang and Stoyanovich, 2017; Das and Lease, 2019), or cross-entropy (Deldjoo et al., 2019).

Several approaches have been adopted to deal with the decreasing value of progressively lower ranking positions. Yang and Stoyanovich (2017) and Zehlike et al. (2017) use simple binomial group distributions (e.g. the fraction of list items produced by the protected group) and average the deviation from the target over successively longer prefixes of the ranking. Sapiezynski et al. (2019) integrate rank discounting into $P(A|L)$ itself, so it is the probability of picking an item by a particular group following the browsing model in the rank discount function.

Provider group fairness can also be assessed over multiple rankings through an exposure metric. Aggregating expected exposure by group (Biega et al., 2018; Diaz et al., 2020) measures whether each producer *group* — rather than individual producer — receives an appropriate level of exposure.

Lastly, pairwise metrics (Beutel et al., 2019) directly implement the concept of disparate mistreatment over pairs of items, extending BPR loss (Rendle et al., 2009) into a fairness construct. A probabilistic ranking is considered fair if the probability of correctly ranking a relevant item (one the user has clicked) over an irrelevant item is independent of the relevant item's group membership. They define two versions of pairwise fairness: intra-group fairness, where the system is equally correct in its rankings of items within each group, and intra-group fairness, where the system correctly orders a relevant item from one group over an irrelevant item from another.

5 Other Stakeholders

While fairness-aware recommendation research to date concentrates largely on consumers and providers, the concept of multistakeholder recommender (see Chapter MULTISTAKEHOLDER) includes others raising fairness concerns that a system should address. Indeed, in some settings, regulatory agencies may be the most important stakeholders for deciding the minimum legal standards with respect to fairness and / or non-discrimination that a system must meet. In other cases, the structure of a platform includes stakeholders who are neither consumers nor providers, but are still impacted by specific parameters of transactions on the platform.

For example, consider a food delivery platform such as UberEats[5] (Abdollahpouri, 2020). This platform uses recommendations to match consumers with restaurants where they might order food to be delivered, and deliveries are made by Uber's drivers. There may be fairness concerns relative to the consumers or providers here, but there may also be fairness concerns over the set of drivers. These individuals do not participate in the recommendation interaction but the recommendations may impact them; for example, a goal might be to ensure that protected groups among the driver population do not receive fewer orders than others or do not receive a disproportionate number of difficult and/or low-tip jobs.

Another example is *subject fairness*, where the goal is to be fair towards the subjects of the items retrieved. In our news recommendation example, this would arise if we wish to ensure that different topics receive fair coverage in the recommended news articles. This fairness concern may break down along traditional lines of discrimination, such as ethnic groups, or it may arise along geographic lines such as urban vs. rural issues and stories. Karako and Manggala (2018) demonstrate an application of a subject fairness constraint in image recommendation. In practice, subject fairness and provider fairness will often be measurable with similar techniques, since both are properties of the items being recommended, but providing one does not necessarily imply the other.

There has been very little published work to date that considers these stakeholders indirectly impacted by recommendation provision, but it is an important direction for future research. The problem of identifying the entities and groups so affected requires significant background data or content analysis.

6 Fairness over Time

A recommendation system operates in an iterated, changing environment, continuously making new decisions, gaining fresh users, and losing established users. Analyzing the fairness of decisions at one point in time can overlook temporal patterns of equity. In the context of predictive policing, Ensign et al. (2018) theoretically demonstrate how feedback loops can result in fair decision-making.

Hashimoto et al. (2018) study feedback loops in production systems. The authors model a population of users iteratively engaging with a system that trains using behavioral data. The model and supporting experiments in the context of predictive typing demonstrate that, over time, machine learning algorithms pay more attention to dominant subgroups of users as they lose under-represented subgroups of users. The authors propose applying techniques from distributionally robust optimization to achieve more balanced performance, resulting in broad user retention. Zhang et al. (2019) extend this work by analyzing the dynamics of fairness in sequential decision-making. Sonboli et al. (2020a) present an adaptive recommendation

[5] http://www.ubereats.com/

approach to multidimensional fairness using probabilistic social choice to control subgroup fairness over time.

Chaney et al. (2018) study the homogenization of recommendations in iterative environments. They find that recommendation systems, especially those based on machine learning, increase the consistency in recommendations across different users but also tend to increase the inequity of exposure across items.

7 Methods for Mitigation

Unfair biases can enter a recommender system at any step of the process, from data collection to evaluation, and there are approaches to mitigating unfairness at each of them. In Sections 3 and 4 we discussed different fairness constructs and evaluation metrics for recommender systems; this section turns to methods for mitigating these biases in data and algorithms (user interface and experience dimensions biases, while important, are out of scope for this chapter). There is no one solution for fairness problems; different applications and objectives will require different techniques.

7.1 Fairness in Data

There has been limited work to date on data-level fairness interventions for recommender systems. Several techniques have appeared in the broader machine learning literature; they may be applicable directly for reducing bias from aspects of item or user representations, but likely need further adaptation to account for the user- and item-level non-independence of the consumption and preference data used to train recommender systems.

Chen et al. (2018) identify different data collection enhancements to cure different kinds of biases resulting from data issues. If there is a group that is under-represented in the training data and therefore receiving lower-quality results, collecting more labeled data from this group can improve the accuracy of their classifications or recommendations. If there is a cluster of users, producers, or items, who experience disparate adverse effects, and that disparity is not explainable by the available features in the data set, then collecting additional features can enhance system fairness.

Other approaches modify the available data in order to reduce unfairness. Feldman et al. (2015) propose to perturb the values of insensitive features X so their distributions are independent of a sensitive attribute A; if insensitive features are no longer correlated with group membership, then the outcomes of machine learning model trained on them will also be uncorrelated with group membership.

Unfortunately, these techniques have not yet — to our knowledge — been adapted to recommendation settings where much of the learning is done from multiple user interactions with an item. Gebru et al. (2018) argue that extensive documentation on motivation for collecting the data, the composition of different components,

collection process, and pre-processing can facilitate design decisions to enhance fairness, a lesson which seems as applicable in recommender systems as in other ML applications.

7.2 Fairness in Ranking Models

The ranking or scoring model itself can also be modified to improve the fairness of results. This usually takes the form of a multi-objective optimization problem that seeks to simultaneously maximize utility (or relevance) and minimize unfairness under some suitable definition discussed in the preceding sections.

Kamishima et al. (2018) does this through *independence*: learning a matrix decomposition that augments the standard regularized reconstruction error loss (see Chapter MATRIX) with an independence term that penalizes correlations between a protected attribute and the system's predicted ratings. They propose multiple non-independence measures, including the difference in means between groups and the mutual information between predicted ratings and sensitive attributes. Independence can also be applied to consumer fairness (Kamishima and Akaho, 2017), with the objective that particular item recommendations are independent of users' sensitive attributes. Yao and Huang (2017) apply a similar regularization approach to minimize disparate rating prediction errors rather than recommendation errors. Beutel et al. (2017) augment pairwise rank loss with a penalty term for disparate rank effectiveness for different groups: this penalty is minimized when the difference in scores between relevant and irrelevant items is uncorrelated with the relevant item's sensitive attribute. It is also possible to directly optimize a learning-to-rank model for equal expected exposure (Diaz et al., 2020).

Constrained optimization approaches such as that of (Biega et al., 2018) produce rankings that maximize user relevance score, subject to constraints on the cumulative attention received by individual items across different users. This can be done through an integer linear programming formulation.

Fair policy learning applies the multi-objective frame to policy learning for a reinforcement learning agent (see Chapter RL); the agent is augmented with a constraint that the expected unfairness (as measured by configurable fairness metric) is less than a threshold δ (Singh and Joachims, 2019).

7.3 Fairness through Re-ranking

Fairness can also be improved by re-ranking lists that are already optimized for relevance, similar to designs used for enhancing diversity (see Chapter DIVERSITY). There is a basic distinction between approaches that treat the problem as a global optimization task, trying to improve the fairness of an entire collection of recommendation lists, and approaches that are focused on a single list.

An example of the global approach is that of Sürer et al. (2018), who suggest a constrained optimization-based method at the post-processing level to enhance fairness for multiple provider groups. This approach is generalizable to multiple constraints, such as a) the inclusion of items from different provider groups, b) ensuring a minimum degree of item diversity for each consumer, and c) avoiding unfairness towards providers from underrepresented populations. The global approach might be suitable in cases where recommendations are all generated at once, such as for a mass personalized email or weekly recommended items list. It could also apply to applications where recommendations are generated in advance and cached.

A more typical approach is to re-rank individual lists as they are generated. Some authors borrow from such approaches as MMR (Carbonell and Goldstein, 1998) and xQUAD (Santos et al., 2010) in the information retrieval literature in proposing greedy list extension methods, where the re-ranked list is built by adding items that, at each point, provide an optimal tradeoff between accuracy and fairness. Geyik et al. (2019) use a greedy approach to produce rankings of job candidates to achieve a desired (non-discriminatory) distribution of candidates' protected attributes, simultaneously optimizing relevance and fairness. Similarly, Modani et al. (2017) re-rank to enhance provider exposure without sacrificing relevance. Zehlike et al. (2017) uses a different approach to the re-ranking problem, using the A-star algorithm to achieve distributional fairness in a ranked list at depth K. The objective of the algorithm is to re-arrange ranked items to meet the desired equity distribution of items from different protected groups while maintaining the rank quality.

Liu et al. (2019) propose a method that incorporates both provider fairness and user preferences for diverse results. Protected groups are promoted by assigning a higher value to items from such groups in the recommendation objective. The promotion of protected groups is tempered by a factor tied to user profile diversity. This method enhances the trade-off between accuracy and fairness by exposing protected group items to users most likely to be interested in them. Sonboli et al. (2020b) extend this approach to multiple protected groups defined by different provider features. Each user can be characterized in terms of the opportunity they provide to promote one or another fairness concern across the different groups.

Reranking can also be applied to consumer fairness. Abdollahpouri (2020) present a re-ranking approach based on the idea of calibration (Steck, 2018) to improve the fairness for user groups with niche interests.

7.4 Fairness through Engineering

Data and algorithmic interventions are not necessarily the best way to achieve some types of fairness. While opportunities for providers to be exposed is zero-sum for a fixed set of recommendation requests (a recommendation slot given to one provider cannot be given to another), consumer recommendation quality is not: solving disparate recommendation utility by decreasing the quality of some users' recommendations does not necessarily improve the quality of others. Instead, identifying the

existence of under-served consumers and reasons their recommendation quality suffers may help in product development, as features which will improve their experience can be prioritized over features that will primarily provide a marginal improvement for users already well-served by the system.

8 Conclusion

Fair recommendation is a relatively new but rapidly growing corner of the recommender systems research literature. The work in this space draws from concerns that have long been of interest to recommender systems researchers, such as those motivating long-tail recommendation and the study of popularity bias, but connects it to the emerging field algorithmic fairness and its roots in the broader literature on fairness and discrimination in general.

Fairness and recommendation are both complex, multifaceted problems, and their combination requires particular clarity on the precise effects and harms to be measured and/or mitigated. In particular, the multisided nature and ranked outputs of many recommender systems complicate the problem of assessing their fairness, as we must identify which stakeholders we are concerned with treating fairly and develop a definition of fairness that applies to specific harms in the context of repeated, ranked outputs, among other challenges. The work of fair recommendation often requires data that is not commonly included with recommender system data sets, particularly when seeking to ensure recommendations are fair with respect to sensitive characteristics users or creators, such as their gender or ethnicity.

This work must also be done with great care and compassion to ensure that users and creators are treated with respect and dignity and to avoid various traps that result in overbroad or ungeneralizable claims. We argue that there is nothing particularly new about these requirements, but that thinking about the fairness of recommendations brings to the surface issues that should be considered in all recommendation research and development.

While there has been much useful work in mapping the problem space and addressing certain types of unfairness, there are many open problems in fair recommendation that need attention. Some of the ones we see include:

- *Extending the concepts and methods of fair recommendation research to additional domains, applications, problem framings, and axes of fairness concerns.* Due to the specific and distinct ways in which social biases and discrimination manifest (Selbst et al., 2019), we cannot assume that findings on one bias translate to another (e.g. findings on race may not apply to ethnicity or geographic location), or that findings on a particular bias in one application will translate to another (e.g. ethnic bias may manifest differently in recommendation vs. NLP classification tasks). Over time, generalizable principles may be discovered and give rise to theories that enable the prediction of particular biases and their manifestations, but at the present time we need to study a wide range of biases and applications to build the knowledge from which such principles may be derived.

- *Deeper study of the development and evolution of biases over time.* Most work — with the exception of fair policy learning and a handful of other studies — focuses on one-shot batch evaluation of recommender algorithms and their fairness. However, recommender behavior is dynamic over time as the system produces recommendations, users respond to them, and it learns from their feedback. This dynamicism means that an initially fair system may become unfair over time if users respond to it in a biased or discriminatory fashion, or that it may move towards a more fair state if users respond well to recommendations that increase overall fairness.
- *Define and study further fairness concerns beyond consumer and provider fairness.* We have identified subject fairness as one additional concern here, but doubt it is the only additional stakeholder whose equity concerns should be considered.
- *Study human desires for and response to fair recommendations.* The first works are beginning to surface in this direction (Smith et al., 2020), and Harambam et al. (2019) explored users' desired features and capabilities for recommendation with concerns that touch on fairness, but at present little is known about what users or content providers expect from a system with respect to its fairness, or how users will respond to fairness-enhancing recommendation interventions.
- *Develop appropriate metrics for recommendation fairness, along with thorough understanding of the requirements and behavior of fairness metrics and best practices for applying them in practical situations.* For example, we believe expected exposure (Diaz et al., 2020) and pairwise fairness (Beutel et al., 2019) are useful frameworks for reasoning about many provider fairness concerns, but there is much work left to do to understand how best to apply and interpret them in offline and online studies.
- *Develop standards and best practices for recommender systems data and model provenance.* Gebru et al. (2018) presented the idea of *datasheets* for data sets, thoroughly and carefully documenting data so downstream users can properly assess its applicability, limitations, and the appropriateness of a proposed use. Harper and Konstan (2015) provide much of this information for the MovieLens data set, but many other data sets do not have such documentation. Mitchell et al. (2019) provide a framework for documenting important properties of trained models, and Yang et al. (2018) present a "nutrition label" for (non-personalized) rankings describing their data sources, ranking principles, and other information. These concepts need to be extended to recommender systems and the complex integrated data sets that drive them. Common practices such as pruning may have deep implications for experiment and recommendation outcomes (Beel and Brunel, 2019), emphasizing the need for careful study of the properties of recommendation data, models, and outputs that should be documented.
- *Engage more deeply with the multidimensional and complex nature of bias.* Most work fair recommendation focuses on single attributes in isolation, often restricting them to binary values. However, the intersection of group memberships often gives rise to particular forms of discrimination and social bias that cannot be explained by any one of the groups alone (Crenshaw, 1989). Some recent work engages with multiple simultaneous axes of discrimination or fairness (Yang et al.,

2020), but multidimensionality does not fully capture the dynamics invoked by the concept of intersectionality (Hoffmann, 2019). Further, many categories are complex, unstable, and socially constructed; Hanna et al. (2020) present a treatment of some of these complexities in an algorithmic setting, but much work remains to respond to that call and make fairness responsive to these realities.

There is a lot of open space for research in fair recommendation, and this work has the potential for significant improvements to the human and societal impact of the systems this book teaches its readers to build and study.

References

Abdollahpouri H (2020) Popularity bias in recommendation: A multi-stakeholder perspective. PhD thesis, University of Colorado Boulder

Ali M, Sapiezynski P, Bogen M, Korolova A, Mislove A, Rieke A (2019) Discrimination through optimization: How Facebook's ad delivery can lead to biased outcomes. Proceedings of the ACM on Human-Computer Interaction 3(CSCW):1–30, DOI 10.1145/3359301

Barocas S, Selbst AD (2016) Big data's disparate impact. California law review 104(3):671, DOI 10.15779/Z38BG31

Barocas S, Hardt M, Narayanan A (2019) Fairness and Machine Learning: Limitations and Opportunities

Beel J, Brunel V (2019) Data pruning in recommender systems research: Best-Practice or malpractice? In: ACM RecSys 2019 Late-Breaking Results

Beutel A, Chen J, Zhao Z, Chi EH (2017) Data decisions and theoretical implications when adversarially learning fair representations. arXiv Preprint 1707.00075

Beutel A, Chi EH, Goodrow C, Chen J, Doshi T, Qian H, Wei L, Wu Y, Heldt L, Zhao Z, Hong L (2019) Fairness in recommendation ranking through pairwise comparisons. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, DOI 10.1145/3292500.3330745

Biega AJ, Gummadi KP, Weikum G (2018) Equity of attention: Amortizing individual fairness in rankings. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, pp 405–414, DOI 10.1145/3209978.3210063

Biega AJ, Diaz F, Ekstrand MD, Kohlmeier S (2020) Overview of the TREC 2019 fair ranking track. In: The Twenty-Eighth Text REtrieval Conference (TREC 2019) Proceedings

Billey A, Haugen M, Hostage J, Sack N, Schiff AL (2016) Report of the PCC ad hoc task group on gender in name authority records. Tech. rep., Program for Cooperative Cataloging

Buolamwini J, Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, PMLR, vol 81, pp 77–91

Burke R (2017) Multisided fairness for recommendation. arXiv Preprint 1707.00093

Burke R, Kontny J, Sonboli N (2018a) Synthetic attribute data for evaluating consumer-side fairness. arXiv Preprint 1809.04199

Burke R, Sonboli N, Ordonez-Gauger A (2018b) Balanced neighborhoods for multi-sided fairness in recommendation. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, vol 81, pp 202–214

Cañamares R, Castells P (2018) Should I follow the crowd?: A probabilistic analysis of the effectiveness of popularity in recommender systems. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, New York, NY, USA, pp 415–424, DOI 10.1145/3209978.3210014

Carbonell J, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, pp 335–336, DOI 10.1145/290941.291025

Celma Ò (2010) Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space. Springer, Berlin, Heidelberg, DOI 10.1007/978-3-642-13287-2

Chaney AJB, Stewart BM, Engelhardt BE (2018) How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In: Proceedings of the 12th ACM Conference on Recommender Systems, ACM, New York, NY, USA, pp 224–232, DOI 10.1145/3240323.3240370

Chen I, Johansson FD, Sontag D (2018) Why is my classifier discriminatory? In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems 31, pp 3539–3550

Crawford K (2017) The trouble with bias. Neural Information Processing Systems

Crenshaw K (1989) Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. The University of Chicago Legal Forum 1989:139–168

Das A, Lease M (2019) A conceptual framework for evaluating fairness in search. arXiv Preprint 1907.09328

Deldjoo Y, Anelli VW, Zamani H, Bellogin A, Di Noia T (2019) Recommender systems fairness evaluation via generalized cross entropy. In: Proceedings of the Workshop on Recommendation in Multi-stakeholder Environments at RecSys '19, CEUR-WS, vol 2440

Diaz F, Mitra B, Ekstrand MD, Biega AJ, Carterette B (2020) Evaluating stochastic rankings with expected exposure. In: Proceedings of the 29th ACM International Conference on Information and Knowledge Management, ACM, DOI 10.1145/3340531.3411962

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R (2012) Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ACM, New York, NY, USA, pp 214–226, DOI 10.1145/2090236.2090255

Ekstrand MD, Kluver D (2021) Exploring author gender in book rating and recommendation. User modeling and user-adapted interaction DOI 10.1007/s11257-020-09284-2

Ekstrand MD, Tian M, Azpiazu IM, Ekstrand JD, Anuyah O, McNeill D, Pera MS (2018a) All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In: Friedler SA, Wilson C (eds) Proceedings of the Conference on Fairness, Accountability, and Transparency (PMLR), New York, NY, USA, PMLR, vol 81, pp 172–186

Ekstrand MD, Tian M, Kazi MRI, Mehrpouyan H, Kluver D (2018b) Exploring author gender in book rating and recommendation. In: Proceedings of the Twelfth ACM Conference on Recommender Systems, ACM, DOI 10.1145/3240323.3240373

Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S (2018) Runaway feedback loops in predictive policing. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, New York, NY, USA, PMLR, vol 81, pp 160–171

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S (2015) Certifying and removing disparate impact. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp 259–268, DOI 10.1145/2783258.2783311

Ferraro A (2019) Music cold-start and long-tail recommendation: bias in deep representations. In: Proceedings of the 13th ACM Conference on Recommender Systems, ACM, New York, NY, USA, pp 586–590, DOI 10.1145/3298689.3347052

Fish B, Bashardoust A, Boyd D, Friedler S, Scheidegger C, Venkatasubramanian S (2019) Gaps in information access in social networks? In: WWW '19: The World Wide Web Conference, ACM, New York, NY, USA, pp 480–490, DOI 10.1145/3308558.3313680

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, Crawford K (2018) Datasheets for datasets. arXiv Preprint 1803.09010

Geyik SC, Ambler S, Kenthapadi K (2019) Fairness-Aware ranking in search & recommendation systems with application to LinkedIn talent search. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA, pp 2221–2231, DOI 10.1145/3292500.3330691

Hamidi F, Scheuerman MK, Branham SM (2018) Gender recognition or gender reductionism?: The social implications of embedded gender recognition systems. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM, p 8, DOI 10.1145/3173574.3173582

Hanna A, Denton E, Smart A, Smith-Loud J (2020) Towards a critical race methodology in algorithmic fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, pp 501–512, DOI 10.1145/3351095.3372826

Harambam J, Bountouridis D, Makhortykh M, van Hoboken J (2019) Designing for the better by taking users into account: a qualitative evaluation of user control mechanisms in (news) recommender systems. In: Proceedings of the 13th ACM

Conference on Recommender Systems, ACM, New York, NY, USA, pp 69–77, DOI 10.1145/3298689.3347014

Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. In: Advances in Neural Information Processing Systems, pp 3315–3323

Harper FM, Konstan JA (2015) The MovieLens datasets: History and context. ACM Transactions on Interactive Intelligent Systems 5(4):19:1–19:19, DOI 10.1145/2827872

Hashimoto T, Srivastava M, Namkoong H, Liang P (2018) Fairness without demographics in repeated loss minimization. In: Dy J, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, Stockholmsmässan, Stockholm Sweden, PMLR, vol 80, pp 1929–1938

Hentschel T, Braun S, Peus CV, Frey D (2014) Wording of advertisements influences women's intention to apply for career opportunities. Academy of Management Proceedings 2014(1):15994, DOI 10.5465/ambpp.2014.15994abstract

Hoffmann AL (2019) Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Information, Communication and Society 22(7):900–915, DOI 10.1080/1369118X.2019.1573912

Hutchinson B, Mitchell M (2019) 50 years of test (un)fairness: Lessons for machine learning. In: FAT 2019: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, pp 49–58, DOI 10.1145/3287560.3287600

Kallus N, Mao X, Zhou A (2019) Assessing algorithmic fairness with unobserved protected class using data combination. arXiv Preprint 1906.00285

Kamishima T, Akaho S (2017) Considerations on recommendation independence for a Find-Good-Items task. In: Workshop on Fairness, Accountability and Transparency in Recommender Systems at RecSys 2017

Kamishima T, Akaho S, Asoh H, Sakuma J (2018) Recommendation independence. In: Friedler SA, Wilson C (eds) Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR, vol 81, pp 187–201

Karako C, Manggala P (2018) Using image fairness representations in Diversity-Based re-ranking for recommendations. In: Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization, ACM, New York, NY, USA, pp 23–28, DOI 10.1145/3213586.3226206

Kaya M, Bridge D, Tintarev N (2020) Ensuring fairness in group recommendations by Rank-Sensitive balancing of relevance. In: Fourteenth ACM Conference on Recommender Systems, ACM, New York, NY, USA, pp 101–110, DOI 10.1145/3383313.3412232

Kearns M, Neel S, Roth A, Wu ZS (2019) An empirical study of rich subgroup fairness for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, pp 100–109, DOI 10.1145/3287560.3287592

Lahoti P, Gummadi KP, Weikum G (2019) ifair: Learning individually fair data representations for algorithmic decision making. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE), pp 1334–1345, DOI 10.1109/ICDE.2019.00121

Liu W, Guo J, Sonboli N, Burke R, Zhang S (2019) Personalized fairness-aware re-ranking for microlending. In: Proceedings of the 13th ACM Conference on Recommender Systems, ACM, DOI 10.1145/3298689.3347016

Mehrotra R, Anderson A, Diaz F, Sharma A, Wallach H, Yilmaz E (2017) Auditing search engines for differential satisfaction across demographics. In: Proceedings of the 26th International Conference on World Wide Web Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp 626–633, DOI 10.1145/3041021.3054197

Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, pp 220–229, DOI 10.1145/3287560.3287596

Mitchell S, Potash E, Barocas S, D'Amour A, Lum K (2020) Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application 8, DOI 10.1146/annurev-statistics-042720-125902

Modani N, Jain D, Soni U, Gupta GK, Agarwal P (2017) Fairness aware recommendations on Behance. In: Advances in Knowledge Discovery and Data Mining, Springer International Publishing, pp 144–155, DOI 10.1007/978-3-319-57529-2_12

Nasr M, Tschantz MC (2020) Bidding strategies with gender nondiscrimination constraints for online ad auctions. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM, New York, NY, USA, pp 337–347, DOI 10.1145/3351095.3375783

Olteanu A, Castillo C, Diaz F, Kıcıman E (2019) Social data: Biases, methodological pitfalls, and ethical boundaries. Frontiers in Big Data 2:13, DOI 10.3389/fdata.2019.00013

Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2009) BPR: Bayesian personalized ranking from implicit feedback. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, Arlington, Virginia, United States, pp 452–461

Santos RLT, Peng J, Macdonald C, Ounis I (2010) Explicit search result diversification through sub-queries. In: ECIR 2010: Advances in Information Retrieval, Springer, LNCS, vol 5993, pp 87–99, DOI 10.1007/978-3-642-12275-0_11

Sapiezynski P, Zeng W, E Robertson R, Mislove A, Wilson C (2019) Quantifying the impact of user attentionon fair group representation in ranked lists. In: Companion Proceedings of The 2019 World Wide Web Conference, ACM, New York, NY, USA, pp 553–562, DOI 10.1145/3308560.3317595

Schedl M (2016) The LFM-1b dataset for music retrieval and recommendation. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, ACM, New York, NY, USA, pp 103–110, DOI 10.1145/2911996.2912004

Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19, ACM, New York, New York, USA, pp 59–68, DOI 10.1145/3287560.3287598

Singh A, Joachims T (2019) Policy learning for fairness in ranking. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems 32, pp 5426–5436

Smith J, Sonbolil N, Fiesler C, Burke R (2020) Exploring user opinions of fairness in recommender systems. In: Fair & Responsible AI Workshop @ CHI 2020

Sonboli N, Burke R, Mattei N, Eskandanian F, Gao T (2020a) "and the winner is...": Dynamic lotteries for multi-group fairness-aware recommendation. 2009.02590

Sonboli N, Eskandanian F, Burke R, Liu W, Mobasher B (2020b) Opportunistic multi-aspect fairness through personalized re-ranking. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, ACM, New York, NY, USA, pp 239–247, DOI 10.1145/3340631.3394846

Steck H (2018) Calibrated recommendations. In: Proceedings of the 12th ACM Conference on Recommender Systems, ACM, pp 154–162, DOI 10.1145/3240323.3240372

Sürer Ö, Burke R, Malthouse EC (2018) Multistakeholder recommendation with provider constraints. In: Proceedings of the 12th ACM Conference on Recommender Systems, ACM, New York, NY, USA, pp 54–62, DOI 10.1145/3240323.3240350

Xiang A, Raji ID (2019) On the legal compatibility of fairness definitions. arXiv Preprint 1912.00761

Yang K, Stoyanovich J (2017) Measuring fairness in ranked outputs. In: Proceedings of the 29th International Conference on Scientific and Statistical Database Management, ACM, New York, NY, USA, Article 22, pp 1–6, DOI 10.1145/3085504.3085526

Yang K, Stoyanovich J, Asudeh A, Howe B, Jagadish HV, Miklau G (2018) A nutritional label for rankings. In: Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18, ACM, New York, New York, USA, pp 1773–1776, DOI 10.1145/3183713.3193568

Yang K, Loftus JR, Stoyanovich J (2020) Causal intersectionality for fair ranking. arXiv Preprint 2006.08688

Yao S, Huang B (2017) Beyond parity: Fairness objectives for collaborative filtering. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in Neural Information Processing Systems 30, pp 2925–2934

Zehlike M, Bonchi F, Castillo C, Hajian S, Megahed M, Baeza-Yates R (2017) FA*IR: A fair top-k ranking algorithm. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, ACM, pp 1569–1578, DOI 10.1145/3132847.3132938

Zhang X, Khaliligarekani M, Tekin C, liu, mingyan (2019) Group retention when using machine learning in sequential decision making: the interplay between user dynamics and fairness. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché Buc F, Fox E, Garnett R (eds) Advances in Neural Information Processing Systems 32, pp 15269–15278