Fair ranking: a critical review, challenges, and future directions

Gourab K. Patro
IIT Kharagpur, India
L3S Research Center, Germany

Lorenzo Porcaro
Universitat Pompeu Fabra
Spain

Laura Mitchell
Competition and Markets Authority
United Kingdom

Qiuyue Zhang
Accenture Plc
United Kingdom

Meike Zehlike
MPI for Software Systems
Zalando Research, Germany

Nikhil Garg
Cornell Tech
United States

ABSTRACT

Ranking, recommendation, and retrieval systems are widely used in online platforms and other societal systems, including e-commerce, media-streaming, admissions, gig platforms, and hiring. In the recent past, a large "fair ranking" research literature has been developed around making these systems fair to the individuals, providers, or content that are being ranked. Most of this literature defines fairness for a single instance of retrieval, or as a simple additive notion for multiple instances of retrievals over time. This work provides a critical overview of this literature, detailing the often context-specific concerns that such approaches miss: the gap between high ranking placements and true provider utility, spillovers and compounding effects over time, induced strategic incentives, and the effect of statistical uncertainty. We then provide a path forward for a more holistic and impact-oriented fair ranking research agenda, including methodological lessons from other fields and the role of the broader stakeholder community in overcoming data bottlenecks and designing effective regulatory environments.

CCS CONCEPTS

• **Information systems → Information retrieval**.

KEYWORDS

Ranking, Recommendation, Fairness, Exposure, Strategic Behaviour, Algorithmic Impact Assessment

ACM Reference Format:
Gourab K. Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair ranking: a critical review, challenges, and future directions. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3531146.3533238

1 INTRODUCTION

Ranking and recommendation systems are ubiquitous across both online marketplaces (e-commerce, gig-economy, multimedia) and other socio-technical systems (admissions or labor platforms), playing a role in which products are bought, who is hired, and what media is consumed. In many of these systems, ranking algorithms form a core aspect of how a large search space is made manageable for *consumers* (employers, buyers, admissions officers, etc). In turn, these algorithms are consequential to the *providers* (sellers, workers, job seekers, content creators, media houses, etc.) who are being ranked.

Much of the initial work on such ranking, recommendation, or retrieval systems[1] (RS) focused on learning to maximize *relevance*—often measured through proxies like clickthrough rate—and thus showing the most relevant items to the consumer, based solely on the consumer's objective [2, 101]. However, like all machine learning techniques, such systems have been found to 'unfairly' favor or discriminate against certain individuals or groups of individuals in various scenarios [8, 34, 47].

Thus, as part of the burgeoning algorithmic fairness literature [110, 115], there have recently been many works on fairness in ranking, recommendation, and constrained allocation more broadly [7, 18, 24, 25, 29, 64, 71, 151, 156, 176, 178]. For example, suppose that the platform is deciding how to rank 10 items on a product search result page, and each item has demographic characteristics (such as those of the seller). Then—in addition to considering each item's relevance—how should the platform rank the items, in a manner that is "fair" to the providers, either on an individual or group level? This question is often considered on an abstract level, independent of the specific ranking context; moreover, the literature primarily focuses on fairness of one instance of the ranking [151, 176–178], or multiple independent instances of rankings with an additive objective across instances [18, 154].

The goals of this paper are to synthesize the current state of the fair ranking and recommendation field, and to lay the agenda for future work. In line with recent papers [82, 148] on both broader fairness and recommendations systems, our view is that the fair ranking literature risks being ineffective for problems faced in real-world ranking and recommendation settings, if it focuses too narrowly on an abstract, static ranking settings. To combat this trend, we identify several pitfalls that have been overlooked in the literature, and should be considered in context-specific ways: toward a broader, long-term view of the fairness implications of a particular ranking system.

Like much of the algorithmic fairness literature, fair ranking mechanisms typically are designed by abstracting away contextual specifics, under a "reducibility" assumption; i.e., many fair

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533238

[1] While we often use "RS" or ranking systems as shorthand, in this work we often mean ranking, recommendation, retrieval, and constrained allocation algorithmic systems more broadly, i.e., systems that select (and potentially order) a subset of providers from a larger available set.

ranking problems of interest can be reduced to a standard problem of ranking, that is a set of items or individuals constrained to a chosen notion of fairness or optimized for a suitable fairness measure (or multiple instances of such ranking over time with simple additive extensions); however, as Selbst et al. [148] elucidate, the abstractions necessary for such a reduction often "render technical interventions ineffective, inaccurate, and sometimes dangerously misguided."

Overview and Contributions. In this work, we outline the many ways in which such a reduction often abstracts away many of the important aspects in the fair ranking context: the gap between position-based metrics and true provider utility, spillovers from one ranking to another across time and products, strategic incentives induced by the system, and the (differential) consequences of ranking noise. Studying fair ranking questions in such a reduced format and ignoring these issues might work in the ideal environment chosen during the problem reduction, but is likely insufficient to bring fairness in a real-world ranking system. For example, a ranking algorithm that does not consider how relevance or consumer discrimination affects outcomes, or how early popularity leads to compounding rewards on many platforms, is unlikely to achieve its fairness desiderata; furthermore, ignoring strategic manipulation (such as Sybil attacks where a provider creates multiple copies of their profile or items) may lead to fairness mechanisms amplifying rather than mitigating inequities on the platform. We believe that these aspects must be tackled by the fair ranking literature, in order for this literature to positively affect practice.

We then overview methodological paths forward to incorporate these aspects into fair ranking research, as part of a broader long-term framework of algorithmic impact assessments—simulations, applied modeling, and data-driven approaches—along with their challenges. Finally, we conclude with a discussion on the broader regulatory, legal, and external audit landscape, necessary to translate the fair ranking literature into systems in practice.

Figure 1 summarizes our paper at a high level.

Outline. Section 2 contains an overview of the fair RS literature. Section 3 presents the aspects of ranking systems that we believe should be most covered by future fair RS work. Section 4 contains the discussion of the paths forward within the broader data and regulatory landscape.

2 OVERVIEW OF FAIR RANKING LITERATURE

Designing effective ranking, recommendation, or retrieval systems (RSs) requires tackling many of the same challenges as to build general machine learning algorithms—with additional challenges stemming from the characteristic that such systems make *comparative* judgments across items; a high position in the ranking is a constrained resource. RSs often employ machine learned models to estimate the *relevance* (or *probability of relevance*) of the items to any search or recommendation query [2, 101]. Historically, while user utility is the broader objective [133], the most popular guiding principle is the *Probability Ranking Principle* [140]: items are ranked in descending order of their probability to be relevant to the user, often estimated through click-through rates. For a broad range of user utility metrics—such as mean average precision [163], mean reciprocal rank [162], and cumulative gain based metrics [84, 85]—this principle in turn maximizes the expected utility of a ranking system for its users [85].

However, not only are more (estimated to be) relevant items typically ranked higher, but also users tend to click more on higher positioned items, even conditioned on relevance. Such a *position bias* [38] means that expected attention (*exposure*) from users decreases significantly while moving from the top rank to the bottom one; for example, users may evaluate items sequentially from the top rank, until they find a satisfactory one. It is thus important for producers to be ranked highly; a small difference in relevance estimation could result in a large difference in expected user attention (for example, see Appendix Table 1). Depending on the ranking context, e.g., ranking products vs. ranking job candidates, high ranking positions directly translate to rewards, or at least increase their likelihood. (However, as we explain in the next section, the gap between exposure and true provider utility is an important one to understand.) Note that despite the recent explorations into multi-sided fairness in online platforms [24, 43, 121, 128, 154], we restrict our discussion to provider fairness which has been studied quite extensively.

Fairness in Rankings. Due to the importance of rankings for providers, and as part of the increased focus on machine learning injustices, there has been much recent interest in fairness and equity for providers rather than just ranking utility for consumers. There are numerous definitions, criteria and evaluation metrics to estimate a system's ability to be *fair* [28, 37, 46, 60, 96, 110, 111, 115, 172]. Given heterogeneous settings, the complex environment in which retrieval systems are developed, and the multitude of stakeholders involved that may have differing moral goals [53] and worldviews [56], there is obviously no universal fairness definition; at a high level, however, many definitions can be classified into whether the objective is to treat similar individuals similarly (*individual fairness*) [44], or if different groups of individuals, defined by certain characteristics such as demographics, should be treated in a similar manner (*group fairness*) [152].

In the following, we overview the concepts and works most relevant for our critiques and the agenda that we advocate. Fairness notions from the domain of classification can—to a certain extent—be adopted to serve in ranking settings. They typically only require additional consideration of the comparative nature of rankings and of how utility is modeled [28]. Compared to relevance-only ranking, adding fairness considerations often leads to the optimization of a multi-objective (or a constrained objective), where the usual utility (or relevance) objective comes along with a fairness constraint or objective focused on the providers [139, 167].

One branch of the literature [7, 29, 64, 176, 178] reasons about probability-based fairness in the top-k ranking positions, which puts the focus onto group fairness. These works commonly provide a minimum (and for some cases also maximum) number or proportion of items/individuals from a protected groups, that are to be distributed evenly across the ranking. The methods do not usually allow later compensation, if the fairness constraints are not met at any of the top-k positions (e.g., by putting more protected than non-protected items to lower positions).

Another set of works [18, 42, 151, 156, 177] assign values (often referred to as *attention* or *exposure* scores) to each ranking position

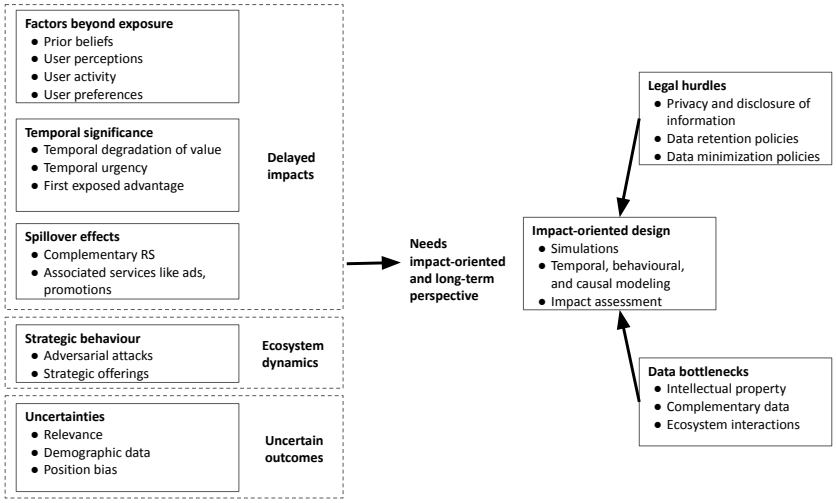

Figure 1: This figure paints a big picture of the paper and succinctly summarizes our position on the field of fairness in retrieval systems, i.e., current fair RS mechanisms often fail to recognize several real-world nuances like delayed impacts, uncertainties in outcomes, ecosystem behaviour (discussed in Section 3); thus we must design fairness interventions in an impact-oriented approach with a holistic and long-term view of RS in mind. In Section 4, we discuss how algorithmic impact assessment can be helpful in this regard. More specifically in Section 4.1, we overview various applied modeling techniques and simulation frameworks which in tandem can be used for impact-oriented studies of fairness in RS. Following this, in Sections 4.2 and 4.3 we briefly discuss various data bottlenecks and legal hurdles which might challenge the efforts towards a holistic view of RS fairness.

based on the expected user attention or click probability. These works argue that the total exposure is a limited resource on any platform (due to position bias), and advocate for fair distribution of exposure to ensure fairness for the providers. In contrast to the former line of work, using exposure as a metric to quantify provider utility has brought up not only group fairness notions [122, 151], but also definitions to enhance individual fairness [18, 23, 151]. Further, in contrast to probability-based methods, these methods balance the *total* exposure across individuals or groups, and thus they do allow compensations in lower positions.

Generally the problem definitions in these works center around a single instance of ranking, i.e., at a particular point in time we are given a set of items or individuals, their sensitive or protected attribute(s) (e.g., race and gender), and their relevance scores; the task is to create a ranking which follows some notion of fairness (like demographic parity or equal opportunity) for the items or individuals, while maximizing the user utility. Some exceptions are Biega et al. [18], Sühr et al. [154] and Sürer et al. [156], that propose to deterministically ensure fairness through equity in amortized exposure, i.e., addition over time or over multiple instances of ranking. In the next section, we argue that both these broad approaches (probability-based, and exposure-based) may be incomplete in many applications, due to their exclusive focus (either directly or indirectly) on ranking positions.

3 PITFALLS OF EXISTING FAIR RANKING MODELS

In this section, we enumerate several crucial aspects of ranking and recommendation systems that substantially influence their fairness properties, but are ignored when considering an abstract fair ranking setting. The left hand side of Figure 1 summarizes this section. We begin in Section 3.1 by noting that exposure (or more generally, equating higher positions with higher utility) often does not translate to provider utility. Section 3.2 discusses spillovers across rankings, either over time, across different rankings on the same user interface, or competition across platforms. Section 3.3 discusses strategic provider responses, and how they may counteract (or worsen) the effects of a fair ranking mechanism. Finally, Section 3.4 illustrates how noise—either in demographic variables or in other aspects—may differentially affect providers within a fair ranking mechanism.

Note that these issues are also present in other aspects of ranking, and in algorithmic fairness literature more generally; in fact, we also discuss if and how such issues have been studied in related settings. However, we believe that the intersection of fairness and ranking challenges amplify these concerns; for example, the naturally comparative aspect of rankings worsens the effects of competitive behavior and differential uncertainties.

Finally, while these pitfalls may not be the only ones, we believe these are the major ones which may cause the failure of proposed

fair ranking frameworks in delivering fair outcomes in several real-world scenarios. In the next section (Section 4), we elaborate on how to tackle these challenges.

3.1 Provider Utility beyond Position-based Exposure

As discussed above, the fair ranking literature often uses *exposure* as a proxy for provider utility[2] [28, 46, 151, 177]. For example, well-known fair ranking mechanisms like *equity of attention* [18] and *fairness of exposure* [151, 177] emphasize fairly allocating exposure among providers. Such works often implicitly assume that exposure is measured solely through a provider's position in the ranking; i.e., each position is assigned a value, independent of context. While such ranking-position-based exposure is often a useful measure of provider utility, such a focus misses context-specific factors due to which higher exposure does not necessarily lead to increased user attention, or that increased user attention may not directly translate to provider utility, as measured through, e.g., sales or long-term satisfaction.

This measurement-construct gap—between exposure as a measurement and provider utility as the construct of interest—is not a challenge unique to fairness-related questions in ranking. For example, not distinguishing between varying levels of attention from users could affect the performance of algorithms designed to maximize sales, as it would affect the predictions of algorithms using exposure to calculate sales probabilities [118] or information diffusion on a social network [10]. However, this gap may be especially important to be considered in a research direction that often seeks algorithmic solutions to inequities stemming from multiple causes, including the actions of other platform participants; for example, much work has analyzed (statistical or taste-based) discrimination on online platforms in which, even conditional on exposure, one type of stakeholders are treated inequitably by other stakeholders (see, e.g., racial discrimination by employers [45, 120]). In such settings, fair-exposure based algorithms may not uniformly or even substantially improve outcomes (we give an example in Appendix Table 2); this was recently underscored by Sühr et al. [155], which found through a user survey that such algorithms' effectiveness substantially depends on context such as job description and candidate profiles.

Another especially relevant contextual factor beyond position is *time*: in fast moving domains like media, items may only be relevant for a short period of time [27, 174]. In such scenarios, the stakeholders (both users and providers) most benefit from immediate exposure. For example, recency is an important aspect of relevance in breaking news [30], job candidates should be shown before vacancies are filled, and restaurants get more orders if recommended during peak hours to nearby customers [11, 174].

More broadly, one should consider *which providers* are being exposed to *which users* and *when*, as the value of a ranking position depends substantially on such match relevance and participant characteristics. Fair ranking models focusing solely on position, and thus oblivious to such context, may not have the desired downstream effects and may fail to deliver on fairness. We illustrate this consequence in an example in Appendix Table 3.

3.2 Spillovers effects: compounding popularity, related items, and competition

While the immediate effect of an item's position in the ranking (e.g., an immediate sale) may be first-order, there are often substantial *spillover* effects or *externalities*, which should be incorporated in fair RS models. Here, we discuss three of such effects: compounding popularity or first-exposed-advantage, spillovers across products and ranking types, and competition effects.

Perhaps the most important spillover is a *compounding popularity* or *first-exposed-advantage*,[3] in which the exposure an item receives during its early stages can significantly affect its long-term popularity [52]. For example, early feedback in terms of clicks, sales, etc. could improve an item's estimated relevance scores, raising its future rankings; there may further be a popularity bias or herding phenomenon in which users are more likely to select an item, if they observe that others have selected it before them [1, 142, 153]. Similarly, as reflected in re-targeting in advertising, user preferences may change with exposure to an item. Thus, past exposure plays a huge role in determining the long-term effects of future exposure; denial of early exposure could risk the viability of small providers [116]. Though one may intuitively think that continuous re-balancing of exposure through fairness-enhancing methods may overcome (or at least reduce) this problem, the real-world-proof is still to be made and early evidence suggests otherwise (see Sühr et al. [155]).

Second, ranking systems—such as product recommendations—are rarely deployed as stand-alone services. They are often accompanied by associated services such as sponsored advertisements [78], similar or complementary item recommendations [138] on individual item pages on e-commerce, media-streaming platforms and other marketplaces [94, 131], non-personalized trending items [16, 39, 132], and other quality endorsements like editor's choice [80]. Due to the presence of these associated services, user attention reaching an item may spill over to other items [98, 136]. For example, complementary items or items similar to an item may receive spillover exposure thereby resulting in increased exposure levels for such items, via 'you may also be interested' or 'items similar to' recommendations, potentially leading to undesirable inequalities even under a fair RS model; we give such an example in Appendix Table 5.

Finally, there are competition and cross-platform spillover effects [49, 92]: users may reach an item, not through the recommendation engine on the platform, but, e.g., via a search engine [83], product or price comparison sites [88], or other platforms like social media [79, 145]. In these instances, the recommendation engine at the user entry-point, e.g., the search engine's recommendation system, will have a downstream effect on the exposure of items on the end site where the items are listed. These spillover effects could be important to analyze when designing potential 'entry-point' recommendation systems. Perhaps more importantly—since a platform

[2]Note that, here we are talking about the utility gained by a provider as a result of getting ranked. Thus provider utility is not same as user utility.

[3]The phrase is used to indicate its similarity to the *first-mover-advantage* phenomenon [89].

does not have control over all the off-platform systems that may influence item exposure on its own platform—one should consider how such external sources affect both the goals and the behavior of a fair RS system. In this regard, the major questions which remain understudied and unanswered at large are: should a fair RS consider the inequities induced via external systems and seek to counteract through interventions or should it ignore these effects for the sake of free market competition?

Together, these spillover effects suggest that fairness in RS (especially in recommendations) should not be modeled in isolation from associated and external services, and must take into account how the recommendations may have downstream consequences over time and space for either the same provider or on other providers. We note that these spillover effects are analogous to the *Ripple Effect trap* as described by Selbst et al. [148], in which harmful effects often stem from the failure of understanding how the introduction of new technologies could alter behaviours and values in existing social systems.

3.3 Strategic Behavior

Current fair ranking mechanisms often fail to consider that the providers themselves could be strategic players who might try to *actively* maximize their utilities [9, 157]. Providers often have an incentive to suitably strategize their offerings, e.g., content creators on media platforms could leave their own area of expertise and try to copy other popular creators or follow the popular trends [13, 14], sellers could perform data poisoning attacks (through fake reviews, views, etc.) on the RS to improve their ranking [180], influencers on social network sites could try to hijack popular trends [31, 66]. Providers can even strategically exploit the deployed fair ranking mechanisms to extract more benefits [57, 125]. Not factoring in such strategic behavior could impact ranking and recommendation systems, and especially the performance of fair ranking mechanisms.

In the following, we overview some examples of strategic behavior and their consequences. As in the measurement-construct gap between exposure and producer utility, strategic behavior as a reaction to ranking models is not just a question of fairness. Numerous works suggest that relevance estimation models are highly vulnerable to various types of adversarial attacks: (1) *shilling attacks*, in which a provider gets associated with a group of users who then add supportive reviews, feedbacks, clicks, etc. to manipulate rankings in favor of the provider [95]; (2) *data poisoning attacks*, where a provider strategically generates malicious data and feeds it into the system through a set of manipulated interactions [97, 180]; or (3) *doppelganger bot attacks*, where a number of fake users or bots are created and then strategically placed in a social network to hijack news feed ranking systems in favor of the malicious party [31, 66, 119].

However, some strategic behavior may specifically exploit characteristics of fair ranking algorithms. For example, fair ranking mechanisms may incentivize *content duplication attacks* [57]. Strategic providers can create duplicates or near-duplicates—possibly hard to automatically identify—of their existing offerings in a ranking system (see the case of Kellogg's Diner [54]). Since certain fair ranking mechanisms may try to ensure benefits for all listed items, providers with more copies of same items stand to gain more benefits [57, 125]. We give such an example in Appendix Table 4. Other 'undesirable' strategic behavior includes the purposeful provision or withholding of information, which may help some participants maximize their ranking; For example, in admissions settings, test-optional admissions policies that aim to be fair to students without test access may inadvertently be susceptible to strategic behavior by students with access but low test scores [102].

Strategic behavior by providers need not always be malicious; rather, it could also represent a sincere effort for improvement (e.g., effort to improve restaurant's quality [104]) or just a change in content offering strategy (e.g., strategic selection of topics for future content production [72, 135]). However, such 'legitimate' strategic behavior may nevertheless affect the efficacy of fair ranking mechanisms over time, as such behavior may affect the relative performance of marketplace participants. For example, Vonderau [161] shows that providers on various content sharing platforms may partly or completely change their content production strategy to cater to the taste of a ranking algorithm (instead of the taste of users). Studies by Chaney et al. [33] and Ben-Porat et al. [13] suggest that ranking mechanisms which are unaware of such behavior could cause homogenization of a platform's item-space and degrade user utility over time; such behavior could also risk the long-term viability and welfare of small-scale providers [116]. Theoretically, Liu et al. [100] extend the strategic classification literature to the ranking setting, to show that such effort (and its differential cost) could have substantial equity implications on the ultimate ranking. Fair ranking mechanisms which seek to equalize exposure affect such incentives, both for desirable and undesirable strategic behavior, and it is necessary to take them into account when designing fair ranking mechanisms for real world settings. Designing fairness mechanisms which can distinguish between such desirable and undesirable behavior may be further challenging (cf. [102]).

Finally, we note that the above discussion—that of strategic behavior of individual providers—does not consider the setting in which the platform—a seemingly neutral player and deployer of a ranking algorithm—also plays the role of a competitive provider (through a subsidiary or partner). Since such providers have access to private platform data and control over their algorithms, they may be able to deploy undetectable strategic manipulations (e.g., Amazon's private label of products on its marketplace [41]) which the other providers are not able to match, leading to an unfair strategy playing field for providers. The design and auditing of ranking algorithms robust to such behavior is an important direction for future work.

3.4 Consequences of Uncertainty

Fairness-aware ranking mechanisms proposed for exposure- and probability-based fairness often assume knowledge of true relevance of providers or items, demographic characteristics on which to remain fair and of the value of each position in the ranking. However, such scores are rarely available in real-world settings. For example, machine-learned models or other statistical techniques used to estimate relevance scores are often uncertain about the relevance of items due to various reasons, for example, biased or noisy feedback, the initial unavailability of data [122, 169], and platform

updates in dynamic settings [130]. While such estimation noise (or bias) is important for all algorithmic ranking or recommendations challenges, it is especially important to consider for fair ranking algorithms, as we illustrate below.

Current fair ranking mechanisms assume the availability of the demographic data of individuals to be ranked. Whilst such assumptions help algorithmic developments for fair ranking, the availability of demographic data can not be taken for granted. Demographic data such as race and gender is often hard to obtain due to reasons like legal prohibitions or privacy concerns on their collection in various domains [4, 21]. To overcome the data gap, platform designers often resort to data-driven inference of demographic information [93], which usually involves huge uncertainty and errors [4]; the use of such uncertain estimates of demographic data in fair ranking mechanisms can cause significant harm to vulnerable groups, and ultimately fail to ensure fairness [65]. Moreover, in dynamic market settings where protected groups of providers or items are often set based on popularity levels, the protected group membership changes over time, thereby adding temporal variations in demographics along with the uncertainty issues [62]. To tackle such variations, Ge et al. [62] propose to use constrained reinforcement learning algorithms which can dynamically adjust the recommendation policy to nevertheless maintain long-term fairness. However, incorporating such demographic uncertainty to broader fair ranking algorithms remains an open question.

Another crucial part of rankings systems is the estimation of position bias [3, 32] which acts as a proxy measure for click-through probability and helps quantify the possible utilities of providers based on their ranks [12]. Fairness-aware ranking mechanisms need these position bias estimates to ensure fair randomized or amortized click-through utility (exposure) for the providers. While these estimates are often assumed to be readily available in most of the recent fair ranking systems works [18, 42, 151], it also has huge uncertainty attached since it heavily depends on the specifics of the user interface. Dynamic and interactive user interfaces [112] used on many platforms, usually go through automatic changes which affects the attention bias (position and vertical bias) based on changes in web-page layout [127]. Furthermore, factors like the presence of attractive summaries and highlighted evidences for relevance—often generated in automated manners—alongside ranking results also differentially affect click-through probabilities over time and across items [86, 175]. Finally, the presence of relevant images, their sizes, text fonts, and other design constraints also play a huge role [68, 103, 166]. Together, as also discussed in Wang et al. [164] and Sapiezynski et al. [144], inaccuracies in position bias estimation and corresponding consequences remain important challenges in fair RS.

Finally, we note that uncertainties, including the above, may be *differential*, affecting some participants more than others, even within the same protected groups. Such differential informativeness, for example, might occur in ranking settings where the platform has more information on some participants (through longer histories, or other access differences) than others [48, 61]. The result of such differential informativeness may cause downstream disparate impact, such as privileging longer-serving providers over newer and smaller ones.

Together, these sources and areas of uncertainty should be an important aspect of future work in fair ranking.

Fair ranking desiderata. What should a comprehensive and long-term view of fairness in RS and its dynamics be composed of? First, the provider utility measure should look beyond mere exposure, and account for user beliefs, perceptions, preferences and effects over time (as discussed in Section 3.1). Second, fair RS works should consider not just immediate impacts but also their spillovers, whether over time for the same item or spillover effects on other items (as discussed in Section 3.2). Third, strategic behavior and systems incentives should also be modeled to anticipate manipulation concerns and their adverse effects (as discussed in Section 3.3). Finally, fair RS mechanisms should incorporate the (potentially differential) effects of estimation noise (as discussed in Section 3.4).

Putting things together, this section illustrated various challenges and downstream effects of developing and deploying algorithms from the fair RS literature. As we discuss in the next section, overcoming these challenges requires both longer-term thinking—beyond the immediate effect of a ranking position—and moving beyond studying general RS settings to modeling and analyzing specific settings and their context-specific dynamics.

4 TOWARDS IMPACT-ORIENTED FAIRNESS IN RANKING AND RECOMMENDER SYSTEMS

In order to avoid the pitfalls discussed in the last section and to design 'truly' fair RS, one must understand and assess the full range and long-term effects of various RS mechanisms. In this regard, we apply recent lessons from and critiques of Algorithmic Impact Assessment (AIA), both within and beyond the FAccT community. Algorithmic Impact Assessment (AIA) can be described as a set of practices and measurements with the purpose of establishing the (direct or indirect) impacts of algorithmic systems, identifying the accountability of those causing harms, and designing effective solutions [113, 137]. More specifically to ranking and recommendation systems, Jannach and Bauer [82] introduces a comprehensive collection of issues related to impact-oriented research in RS. There are two broad lessons from this literature, that we explain and apply to the design of fair RS, in a manner that involves integrated effort from different actors and a comprehensive view of their effects.

First, as discussed by Vecchione et al. [160], a key point when assessing or auditing algorithmic systems is to move *beyond discrete moments of decision making*, i.e., to understand how those decision-points affect the long-run system evolution; this point is particularly true for fairness interventions in ranking and recommender systems, as discussed in Section 3. Jannach and Bauer [82] also highlight the limitations and unsuitability of traditional research in RS, which focused solely on accurately predicting user ratings for items ("leaderboard chasing") or optimizing click-through rates. Thus, in Section 4.1, we begin with a discussion of methodologies that can be used to study such long-run effects of fair RS mechanisms, that have been used to study other questions in RS fields – mainly, simulation and applied modeling. We detail not only the useful frameworks but also potential limitations and challenges when studying fairness-specific questions.

Second, a key aspect of effective assessments is the participation of every suitable stakeholder, including systems developers, affected communities, external experts, and public agencies; otherwise, a danger is that the research community focuses on impacts most measurable by its preferred methods and ignores others [113]. However, there are bottlenecks to such holistic work, especially for RS used in private or sensitive contexts. We discuss data availability challenges in Section 4.2. Then, in Section 4.3, we overview various regulatory frameworks – along with their limitations – designed to govern RS or algorithmic systems in general, and hold them accountable. Researchers should contribute to tackling these challenges as well.

4.1 Simulation and Applied Modeling to Study Long-term Effects and Context-specific Dynamics

Many of the challenges discussed in Section 3 are regarding impacts that do not appear in the short-term, immediately after a given ranking; for example, it may take time for strategic agents to respond to a ranking systems. These long-term impacts are difficult to capture without considering a specific context, or with solely relying on "traditional" metrics that assess instantaneous precision-fairness trade-offs.

Outside of fair ranking, the recommendations literature has investigated such long-term and indirect effects using *simulation and applied modeling* methods, motivated for example by the observation that offline (and commonly, precision-driven) recommendation experiments are not always predictive of long-term simulation or online A/B testing outcomes [20, 67, 91]. However, surprisingly, such an approach has been relatively rare in the fair rankings and recommendations literature; to spur such work, here we overview various simulation and modeling tools along that are advantageous in our context.

First, **simulations** have already been used in the past to demonstrate long-term effects of recommender systems and search engines—although unrelated to fairness, in ways that static precision-based analyses can not. Examples are the demonstration of the *performance paradox* (users' higher reliance on recommendations may lead to lower RS performance accuracy and discovery) by Zhang et al. [181], the study of *homogenization* effects on RS users by Chaney et al. [33], a study on the emergence of *filter bubbles* [123] in collaborative filtering recommendation systems and its impacts by Aridor et al. [5], the evaluation of reinforcement learning to rank for search engines by Hu et al. [81], and a study on *popularity bias* in search engines by Fortunato et al. [55]. All relied on context-specific simulations of RS. Many other works also leverage simulations [11, 22, 40, 51, 76, 107, 129, 130, 171] to study various dynamics in recommender systems. In summary, these works illustrate how simulation-based environments can help in *(i)* studying various hypothesized relationships between the usage of systems and individual and collective behavior and effects, *(ii)* detecting new forms of relationships, and *(iii)* replicating results obtained in empirical studies.

Given the usefulness of simulations, many simulation frameworks have been developed to study various fairness approaches for information retrieval systems; just to mention a few: MARS-Gym [143], ML-fairness-gym [40], Accordion [109], RecLab [91], RecSim NG [117], SIREN [22], T-RECS [106], RecoGym [141], AESim [59], Virtual-Taobao [150].

Note however, that the simulated environments are created under certain assumptions on the interactions between the stakeholders and the system, which may not always hold in real-world. As emphasized by Friedler et al. [56], it is important to question how different value assumptions may be influential on the simulated environments, and which worldviews have been modeled while developing such frameworks. On a positive note, simulation frameworks can be designed to be flexible enough to give freedom in (de)selecting or changing the fundamental value assumptions in fair RS; for example RecoGym [141] and MARS-Gym [143] provide freedom in setting various types of user behaviours and interactions with the system. This flexibility allows impact and efficacy assessment under different ethical scenarios, and the study of fair RS mechanisms under various delayed effects and user biases (as discussed in Sections 3.1 and 3.2) – we believe that leveraging such simulation frameworks is an important path forward to studying the various effects discussed above in a context-specific manner.

Second, various **temporal, behavioural and causal models** have traditionally been used to formally define, understand and study complex dynamical systems in fields like social networks [50, 73, 74], game theory and economics [6, 26], machine learning [70, 170], and epidemiology [69]. These models often rely on real-world observations of individual behaviour, extract broader insights, and then try to formally represent both individual and system dynamics through mathematical modeling. While the simulation frameworks can function as technical tools to study RS dynamics, suitable temporal, behavioural and causal models can be integrated within the simulation to ensure that the eco-system parametrization, stakeholder behaviour and system pay-offs are representative of the real-world. A good example: Radinsky et al. [134] try to improve search engine performance with the use of suitable behavioural and temporal models in their framework. Similarly, simulation frameworks with suitable applied modeling can be used to design and evaluate fair RS mechanisms which can withstand strategic user behaviour and other temporal environment variations. Causal models can be utilized to study the impact of fair RS [147, 149, 165] in presence or absence of uncertainties and various associated services. Applied modeling tools are further an effective way to study strategic concerns in ranking, along with their fairness implications [100].

Even though simulations along with applied modeling may not exactly mirror the real world effects of fair RS, they could give enough of a basis to highlight likely risks, which could then be taken into account while designing and optimizing fair RS mechanisms. They also bring an opportunity to model the effects of proposed fairness interventions, so that their long-term and indirect effects can be better understood and compared.

However, these approaches would further benefit from availability of certain data and the resolution of related legal bottlenecks. For example, studies on spillover effects can not proceed without the data on complementary and associated services. These data and legal bottlenecks might have also contributed to the fact that there are very few works exploring this direction, and out of the limited works, some are limited to either theoretical analysis [13, 116] or

simulations with assumed parametrizations [62, 168, 181] in absence of complementary data.[4] We discuss these bottlenecks in Section 4.2 and Section 4.3.

4.2 Data Bottlenecks

A major challenge faced by researchers outside industry working on long-term comprehensive evaluations of fair RS is the unavailability of suitable data.

The traditional RS datasets [17, 75, 108, 124] that often used in the literature were collected in times when goals like accuracy or click-through rates and so may not be a good fit for today's impact-oriented research [82]. For example, a set of user-item ratings data such as the canonical MovieLens dataset [75] may not capture how a user may value the item differently at different points in time or how a user's preferences evolve over time, or the user's or item's associated demographics. Similarly, such data gives little insight into fake reviews or ratings [77, 97, 105, 180], or other strategic manipulations as discussed above. More broadly, such datasets do not include vital information such interface design changes that may have a behavioural impact on user choice (as discussed in Section 3.4), and associated services like complementary recommender systems or embedded advertisement blocks (as elaborated in Section 3.1) that work alongside the one being audited, the type and time of provider interactions and changes in their behaviour. Such missing components of standard ranking and recommendation system datasets are a major bottleneck to studying the questions from Section 3.

On the other hand, the flourishing of the algorithmic fairness literature have contributed to the spread of several experimental datasets covering a wide range of scenarios such as school admission, credit score, house listings, news articles, and much more (see [110, 179] for a list of datasets used in fair ranking and ML research). Datasets such as *COMPAS* or the *German Credit* datasets, originally classification tasks, have been adapted to ranking settings. A major issue related to the use of these datasets in fair ranking research is that they are often far from the contexts in which fair ranking algorithms would be used. While potentially useful in the advancing the conceptual state-of-the-art in algorithmic fairness research, reliance on such datasets may raise significant concerns to the ecological validity of such research. Therefore, a more detailed analysis on the use and characteristics of such datasets is a much needed work to address in future, similarly to what has been done in the context of Computer Vision research [90, 114, 146].

Here, we detail the characteristics that a RS dataset would need to be suitable for impact-oriented fairness analysis, in addition to the traditional indicators of user preference or experience (precision or click through rates). One recurring theme is that ranking and recommendation systems operate within a broader socio-technical environment (that they themselves shape), and existing datasets do not allow researchers to understand this broader environment and the underlying dynamics.[5]

(1) Most easily, it would be useful to complement existing datasets with past data on the same platform, such as user-provider interactions and their behaviour; on RS's associated services and related rankings; on other contextual details such as user interface, page layout and design; and on past results from rankings, such as whether the user selected a custom sorting criteria like date or price instead of platform's default ranking criteria, whether the user was redirected to a product from an external or affiliate link, and whether the user's behaviour follows the platform's guidelines. Such complementary data would allow understand how the broader environment affects and is affected by a fair ranking algorithm.
(2) More broadly, a move from static datasets to temporal datasets – with timestamps on ratings and displayed recommendations/ratings – would allow finding temporal variations in RS and its stakeholders. It would further allow studying fairness beyond demographic characteristics, such as that related to new providers. For example, as discussed in Section 3.2, higher ranked results can often lead to increased user attention and conversion rates [38], i.e., results initially ranked higher could then have a greater chance of being ranked highly in subsequent rankings. Since such biased feedback could easily creep into temporal datasets, one must factor this in their RS impact analysis (e.g., an unbiased learning method by Joachims et al. [87] in presence of biased feedback). Studying such dynamics and their fairness implications in the real world requires observing such interactions.
(3) Finally, as discussed in Section 3.4, a key aspect of fairness in rankings is uncertainty, especially differential uncertainty. While some datasets may allow researchers to infer certain components of recommendation system uncertainty (such as by numbers of ratings for a provider), other uncertainties are hidden. External to such companies, it is unclear how to best reflect the correctness of provided user attributes (such as race and gender so as to avoid uncertainties in a platform's compliance to fairness), the genuineness of ratings and reviews (so as to account for manipulations in fair RS analysis) [158, 173]) when feedback is given, and other model uncertainties. While it may be difficult for companies to quantify their uncertainties when releasing datasets, one beneficial step would be to release more information on the origin of the data, i.e., dataset datasheets as described by Gebru et al. [63].

Unfortunately, as might be expected, there are several challenges to such comprehensive datasets.

The most important challenges are from the legal domain, which might even affect researchers and developers within a company. For example, the data minimization principle in GDPR [159] could restrict platforms to collect sensitive information like gender or race, thereby indirectly closing the doors for the implementation of fairness interventions, and inferred attributes would contain huge uncertainty which may render fairness interventions useless

[4] Note that a few recent works look into long-term assessment of fair machine learning [40, 99, 182], which we overlook so as not to divert from the primary focus of our discussion.

[5] We note that while *more* data is not always better (e.g., see the case of NLP models discussed by Bender et al. [15])—we believe that a certain level of *completeness and richness of data* is required to perform more comprehensive and long-term impact analysis.

(as discussed in Section 3.4). In fact, a study by Biega et al. [19] finds that the performance might not substantially decrease due to data minimization, but it might disparately impact different users. Additional legal principles which might present challenges are other privacy regulations, data retention policies, intellectual property rights of platforms, etc. We discuss these challenges in the next section.

Furthermore, while a comprehensive and long-term view on fair RS may be of huge societal need and expectation, the creation of suitable datasets and their availability to external researchers heavily rely on the interests of platform owners. Such external access, even if restricted in various ways, is an important aspect of regulation and auditing.

We now turn to discussing such legal and regulatory concerns.

4.3 Legal Bottlenecks

In the previous section we discussed issues of missing data and the challenges to obtain necessary information due to platform interests and legal regulations on privacy. Regulations and other legal interventions by governments are helpful in some aspects of ensuring external audits, while hindering fair ranking and recommendation in other contexts. Legal provisions will vary across jurisdictions, causing different challenges in data access and algorithmic disclosure depending on the location of: the data requested, the users of platforms that implement RS's, the individuals impacted by the rankings, and the researchers seeking access to RS information. For example, data protection laws may potentially restrict access to data located in the EU, for non-EU based researchers or vice versa.

In this section we give an overview of legal hurdles that prevent researchers of fair RS from assessing the impact of their methods, along with information on specific laws and guidelines that can be used as a starting point for discussions to shape a more robust set of legal provisions for long term fair RS.

There are existing laws/guidance that could be applied to long term fairness in RS. But the wording of some of these laws/guidance leaves them open to interpretation, such that a platform could reasonably argue that it is fulfilling its obligations under the guidance, without taking into account long term fairness in RS. The European Commission Ethics Guidelines for Trustworthy AI [126] state that a system should be tested and validated to ensure it is working as intended throughout its entire life cycle, both during development and after deployment. The guidelines list fairness as well as societal well-being as a requirement of trustworthy AI. However, if the word "intended" is interpreted narrowly, as point in time and in isolation from the dynamic and interconnected nature of recommendations, platforms could demonstrate that their systems are working as "intended," considering both fairness and societal impact—even if in practice the platform may not be evaluating for long-term fairness or modelling various spillover effects.

In addition, the European Commission Guidelines on Ranking Transparency [35] reflect hesitancy that platforms have to be fully transparent on the details of their ranking; they recognise that providers are "not required to disclose algorithms or any information that, with reasonable certainty, would result in the enabling of deception of consumers or consumer harm through the manipulation of search results." This privacy-transparency trade-off may cause the problem of missing data for algorithmic impact assessments to continue.

On the other hand, there is a push from regulators to make data from algorithmic systems available—if not to the general public, at least to independent third party auditors—to mitigate conflicts of interest when platforms audit their own systems. In the US, the FTC's Algorithmic Accountability Act [58] provides that if reasonably possible, impact assessments are to be performed in consultation with external third parties, including independent auditors and technology experts. However, the EU harmonised rules for AI [36] acknowledge that given the early phase of the regulatory intervention and the fact the AI sector is very innovative, expertise for auditing is only now being accumulated.

In the absence of underlying data and full knowledge of the ranking algorithm, researchers could still adopt a forward looking approach of implementing simulations, based on what they do know about the ranking, to help predict the longer term effects of a ranking algorithm (as already explained in Section 4.1). It remains to be seen however, whether the advised disclosure of "meaningful explanations" of the main parameters of ranking algorithms—referred to in the European Commission Guidelines on Ranking Transparency [35]—provide enough information upon which to base an evaluation of the long term fairness of the RS. There is also uncertainty over whether these meaningful explanations reduce sufficiently the impact of information asymmetry between users of the platform, and the platform itself, particularly where the platform both controls the RS, and includes its own items to be eligible in ranking results, alongside those of third party providers. Further consideration also needs to be given to the timing of the release of the explanations when an RS method is updated, to give stakeholders sufficient opportunity to challenge reliance on these parameters, from a long term fairness perspective, pre-implementation of the RS update.

Applying laws to, or developing laws for, long term fairness scenarios in RS is in its infancy. Those involved in shaping this legal framework should consider for long term fairness evaluation purposes: data access for different stakeholders, timings for this access, and level of detail that needs to be given; as well as providing actionable guidance on a platform's responsibility for developing RS with long term fairness goals in mind.

5 CONCLUSION

In this paper we provided a critical overview of the current state of research on fairness in ranking, recommendations, and retrieval systems, and especially the aspects often abstracted away in existing research. Much of the existing research has focused on instance-based, static fairness definitions that are prone to oversimplifying real-world ranking systems and their environments. Such a focus may do more harm than good and result in 'fair-washing,' if those methods are deployed without continuous critical investigation on their outcomes. Guidelines and methods to consider the effects of the entire ranking system through its life cycle, including effects from interactions with the outside world, are urgently needed.

We discussed various aspects beyond the actual ordering of items that affect rankings, such as spillover effects, temporal variations,

and varying user characteristics ranging from their levels of activity. We further examined the effects of strategic behaviors and uncertainties in an RS. These effects play an important role for the successful creation and assessment of fair rankings, and yet they are rarely considered in state-of-the-art fair ranking research. Finally, we proposed next steps to overcome these research gaps. As a promising first step we have identified simulations frameworks and applied-modeling methods, which can reflect the complexity of ranking systems and their environments. However, in order to create meaningful impact analysis, concerns around datasets for fair ranking research, certain data bottlenecks and legal hurdles are yet to be resolved.

Our analysis concerning existing research gaps is of course by no means exhaustive, and many other issues of high complexity remain to be discussed. In this paper, we focused on fair ranking methods that try to enhance fairness for a single side of stakeholders, mostly the individuals being ranked, or the providers of items that are ranked. Research that is concerned with multi-stakeholder problems has recently started to emerge—finding, for example, that fairness objectives for providers and consumers in conflict to each other.

Similarly, we also did not explicitly discuss ranking platforms as two-sided markets, in which both sides may receive rankings for the other side. While it is a promising direction with a vast corpus of economic research on the topic, it is important to understand that (1) not all ranking platforms and their environments are two-sided in a literal sense: e.g., Amazon is a platform and a provider at the same time; and (2) depending on what is happening on the platform, different justice frameworks have to be applied: e.g., school choice, LinkedIn, and Amazon can all be seen as two-sided markets in a broader sense, but they need very different approaches when it comes to the question on what it means for them to be fair. Depending on whether people or products are ranked, one might expect different user bias manifestations, as well as different requirements on data privacy and minimization policies. These differences have to be taken into account when designing fair ranking methods.

Additionally, we did not address the challenge to disentangle the origins of bias, namely whether bias comes from design choices of the platform (e.g., rating systems being implemented as 5-star systems vs. a single like button without the possibility to dislike something), or from the users themselves (e.g., gender bias in evaluation of women).

Finally, we note that, to the best of our knowledge, all known definitions of fairness in ranking are drawn from an understanding of fairness as distributive justice: (limited) *primary goods*—these are goods essential for a person's life, such as housing, access to job opportunities, health care, etc.—are to be distributed fairly across a set of individuals. Fair ranking definitions of this kind may be a good fit for hiring or admissions, because we distribute a limited number of primary goods, namely jobs and education, among a set of individuals. However, fairness definitions based on the distributive justice framework may not make sense in other scenarios. For instance, e-commerce platforms may not qualify for properties of distributive justice, because they lack the aspect to distribute *primary* goods: e-commerce settings, e.g., whether a single item is sold, may not qualify as immediately life-changing.

Overall, we conclude that there is still a long way ahead of us; many more aspects from the ranking systems' universe have to be considered before we achieve substantive and robust algorithmic justice in rankings, recommendations, and retrieval systems.

ACKNOWLEDGMENTS

The authors would like to thank Francesco Fabbri, Jessie Finocchiaro, Faidra Monachou, Ignacio Rios, and Ana-Andreea Stoica for helpful comments. This project has been a part of the MD4SG working group on Bias, Discrimination, and Fairness. This research has received funding under European Research Council (ERC) Marie Sklodowska-Curie grant (agreement no. 860630) for the project NoBIAS, Spanish Ministerio de Ciencia, Innovacion y Universidades (MCIU) and the Agencia Estatal de Investigacion (AEI)-PID2019-111403GB-I00/AEI/10.13039/501100011033. G. K Patro acknowledges the support by TCS Research fellowship.

REFERENCES

[1] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 42–46. https://doi.org/10.1145/3109859.3109912

[2] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749. https://doi.org/10.1109/TKDE.2005.99

[3] Aman Agarwal, Ivan Zaitsev, Xuanhui Wang, Cheng Li, Marc Najork, and Thorsten Joachims. 2019. Estimating position bias without intrusive interventions. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 474–482. https://doi.org/10.1145/3289600.3291017

[4] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 249–260. https://doi.org/10.1145/3442188.3445888

[5] Guy Aridor, Duarte Goncalves, and Shan Sikdar. 2020. Deconstructing the Filter Bubble: User Decision-Making and Recommender Systems. In *RecSys 2020 - 14th ACM Conference on Recommender Systems*. 82–91. https://doi.org/10.1145/3383313.3412246

[6] Dan Ariely and Simon Jones. 2008. *Predictably irrational*. Harper Audio New York, NY.

[7] Abolfazl Asudeh, H.V. Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing fair ranking schemes. In *Proceedings of the 2019 International Conference on Management of Data*. 1259–1276. https://doi.org/10.1145/3299869.3300079

[8] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61. https://doi.org/10.1145/3209581

[9] Gal Bahar, Rann Smorodinsky, and Moshe Tennenholtz. 2016. Economic recommendation systems. In *EC '16: Proceedings of the 2016 ACM Conference on Economics and Computation*. https://doi.org/10.1145/2940716.2940719

[10] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*. 519–528. https://doi.org/10.1145/2187836.2187907

[11] Ashmi Banerjee, Gourab K. Patro, Linus W. Dietz, and Abhijnan Chakraborty. 2020. Analyzing 'Near Me' Services: Potential for Exposure Bias in Location-based Retrieval. *2020 IEEE International Conference on Big Data (Big Data)* (2020), 3642–3651. https://doi.org/10.1109/BigData50022.2020.9378476

[12] Judit Bar-Ilan, Kevin Keenoy, Mark Levene, and Eti Yaari. 2009. Presentation bias is significant in determining user preference for search results—A user study. *Journal of the American Society for Information Science and Technology* 60, 1 (2009), 135–149. https://doi.org/10.1002/asi.20941

[13] Omer Ben-Porat, Itay Rosenberg, and Moshe Tennenholtz. 2020. Content Provider Dynamics and Coordination in Recommendation Ecosystems. *Advances in Neural Information Processing Systems* 33 (2020).

[14] Omer Ben-Porat and Moshe Tennenholtz. 2018. A game-theoretic approach to recommendation systems with strategic content providers. *Advances in Neural Information Processing Systems 31* (2018).

[15] Emily M. Bender, Angelina McMillan-Major, Timnit Gebru, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: can language models be too big?. In *FAccT 2021 - Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*, Vol. 1. 271–278. https://doi.org/10.1145/3442188.3445922

[16] James Benhardus and Jugal Kalita. 2013. Streaming trend detection in Twitter. *International Journal of Web Based Communities* 9, 1 (2013), 122–139. https://doi.org/10.1504/IJWBC.2013.051298

[17] James Bennett, Stan Lanning, et al. 2007. The netflix prize. In *Proceedings of KDD cup and workshop*, Vol. 2007. New York, NY, USA., 35.

[18] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international ACM SIGIR Conference on Research and Development in Information Retrieval*. 405–414. https://doi.org/10.1145/3209978.3210063

[19] Asia J Biega, Peter Potash, Hal Daumé, Fernando Diaz, and Michèle Finck. 2020. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 399–408. https://doi.org/10.1145/3397271.3401034

[20] Anand V Bodapati. 2008. Recommendation systems with purchase data. *Journal of marketing research* 45, 1 (2008), 77–93.

[21] Miranda Bogen, Aaron Rieke, and Shazeda Ahmed. 2020. Awareness in practice: tensions in access to sensitive attribute data for antidiscrimination. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 492–500. https://doi.org/10.1145/3351095.3372877

[22] Dimitrios Bountouridis, Jaron Harambam, Mykola Makhortykh, Mónica Marrero, Nava Tintarev, and Claudia Hauff. 2019. SIREN: A simulation framework for understanding the effects of recommender systems in online news environments. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. 150–159. https://doi.org/10.1145/3287560.3287583

[23] Amanda Bower, Hamid Eftekhari, Mikhail Yurochkin, and Yuekai Sun. 2021. Individually Fair Ranking. *ICLR* (2021).

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

[25] William Cai, Johann Gaebler, Nikhil Garg, and Sharad Goel. 2020. Fair allocation through selective information acquisition. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 22–28.

[26] Colin F Camerer. 2003. Behavioural studies of strategic thinking in games. *Trends in cognitive sciences* 7, 5 (2003), 225–231.

[27] Pedro G Campos, Fernando Díez, and Iván Cantador. 2014. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction* 24, 1 (2014), 67–119.

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

[29] L Elisa Celis, Damian Straszak, and Nisheeth K Vishnoi. 2018. Ranking with Fairness Constraints. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, Vol. 107. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 28.

[30] Abhijnan Chakraborty, Saptarshi Ghosh, Niloy Ganguly, and Krishna P Gummadi. 2017. Optimizing the recency-relevancy trade-off in online news recommendations. In *Proceedings of the 26th International Conference on World Wide Web*. 837–846.

[31] Abhijnan Chakraborty, Gourab K Patro, Niloy Ganguly, Krishna P Gummadi, and Patrick Loiseau. 2019. Equality of voice: Towards fair representation in crowdsourced top-k recommendations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 129–138.

[32] Praveen Chandar and Ben Carterette. 2018. Estimating clickthrough bias in the cascade model. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1587–1590.

[33] Allison J B Chaney, Brandon M Stewart, and Barbara E Engelhardt. 2018. How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility. In *RecSys '18 Proceedings of the 12th ACM Conference on Recommender Systems*. https://doi.org/10.1145/3240323.3240370 arXiv:1710.11214

[34] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2020. Bias and Debias in Recommender System: A Survey and Future Directions. *arXiv preprint arXiv:2010.03240* (2020).

[35] European Commission. 2020. *Guidelines on ranking transparency pursuant to Regulation (EU) 2019/1150 of the European Parliament and of the Council*. Retrieved August 18, 2021 from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020XC1208(01)&rid=2

[36] European Commission. 2021. *Proposal for a Regulation laying down harmonised rules on artificial intelligence*. Retrieved June 03, 2021 from https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence

[37] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[38] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. 2008. An experimental comparison of click position-bias models. In *Proceedings of the 2008 international conference on web search and data mining*. 87–94.

[39] Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*. 39–46.

[40] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. 2020. Fairness is not static: Deeper understanding of long term fairness via simulation studies. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 525–534. https://doi.org/10.1145/3351095.3372878

[41] Abhisek Dash, Abhijnan Chakraborty, Saptarshi Ghosh, Animesh Mukherjee, and Krishna P Gummadi. 2021. When the Umpire is also a Player: Bias in Private Label Product Recommendations on E-commerce Marketplaces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 873–884.

[42] Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 275–284.

[43] Virginie Do, Sam Corbett-Davies, Jamal Atif, and Nicolas Usunier. 2021. Two-sided fairness in rankings via Lorenz dominance. *Advances in Neural Information Processing Systems* 34 (2021).

[44] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[45] Benjamin Edelman, Michael Luca, and Dan Svirsky. 2017. Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics* 9, 2 (2017), 1–22.

[46] Michael D. Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and Discrimination in Retrieval and Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1403–1404. https://doi.org/10.1145/3331184.3331380

[47] Michael D Ekstrand, Mucun Tian, Ion Madrazo Azpiazu, Jennifer D Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All the cool kids, how do they fit in?: Popularity and demographic biases in recommender evaluation and effectiveness. In *Conference on Fairness, Accountability and Transparency*. PMLR, 172–186.

[48] Vitalii Emelianov, Nicolas Gast, Krishna P Gummadi, and Patrick Loiseau. 2020. On fair selection in the presence of implicit variance. In *Proceedings of the 21st ACM Conference on Economics and Computation*. 649–675.

[49] Ayman Farahat and Tarun Bhatia. 2016. App Installs on iOS and Android: Cross Platform Spillover. *Available at SSRN 2759557* (2016).

[50] Mehrdad Farajtabar, Yichen Wang, Manuel Gomez-Rodriguez, Shuang Li, Hongyuan Zha, and Le Song. 2017. Coevolve: A joint point process model for information diffusion and network evolution. *The Journal of Machine Learning Research* 18, 1 (2017), 1305–1353.

[51] Andres Ferraro, Dietmar Jannach, and Xavier Serra. 2020. Exploring Longitudinal Effects of Session-based Recommendations. In *RecSys 2020 - 14th ACM Conference on Recommender Systems*. 1–8. https://doi.org/10.1145/3383313.3412213 arXiv:2008.07226

[52] Flavio Figueiredo, Jussara M Almeida, Marcos André Gonçalves, and Fabricio Benevenuto. 2014. On the dynamics of social media popularity: A YouTube case study. *ACM Transactions on Internet Technology (TOIT)* 14, 4 (2014), 1–23.

[53] Jessie Finocchiaro, Roland Maio, Faidra Monachou, Gourab K Patro, Manish Raghavan, Ana-Andreea Stoica, and Stratis Tsirtsis. 2021. Bridging Machine Learning and Mechanism Design towards Algorithmic Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 489–503.

[54] Luke Fortney. 2022. *Why Is Kellogg's Diner Selling Food Under 18 Different Restaurant Names on Delivery Apps?* Retrieved April 19, 2022 from https://ny.eater.com/2022/4/14/23016676/kelloggs-diner-ghost-kitchens-virtual-brands-williamsburg-nyc

[55] Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. 2006. Topical interests and the mitigation of search engine bias. *Proceedings of the national academy of sciences* 103, 34 (2006), 12684–12689.

[56] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making. *Commun. ACM* (2021).

[57] Maik Fröbe, Jan Philipp Bittner, Martin Potthast, and Matthias Hagen. 2020. The effect of content-equivalent near-duplicates on the evaluation of search engines. In *European Conference on Information Retrieval*. Springer, 12–19.

[58] FTC. 2019. *H.R.2231 - Algorithmic Accountability Act of 2019*. Retrieved August 18, 2021 from https://www.congress.gov/bill/116th-congress/house-bill/2231/text

[59] Yongqing Gao, Guangda Huzhang, Weijie Shen, Yawen Liu, Wen-Ji Zhou, Qing Da, and Yang Yu. 2021. Imitate TheWorld: A Search Engine Simulation Platform. *arXiv preprint arXiv:2107.07693* (2021).

[60] David García-Soriano and Francesco Bonchi. 2021. Maxmin-fair ranking: individual fairness under group-fairness constraints. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 436–446.

[61] Nikhil Garg, Hannah Li, and Faidra Monachou. 2021. Standardized tests and affirmative action: The role of bias and variance. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 261–261.

[62] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, et al. 2021. Towards Long-term Fairness in Recommendation. *WSDM* (2021).

[63] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for Datasets. (2018). https://doi.org/10.1145/3458723 arXiv:1803.09010

[64] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2221–2231.

[65] Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When Fair Ranking Meets Uncertain Inference. (2021).

[66] Oana Goga, Giridhari Venkatadri, and Krishna P Gummadi. 2015. The doppelganger bot attack: Exploring identity impersonation in online social networks. In *Proceedings of the 2015 internet measurement conference*. 141–153.

[67] Carlos A Gomez-Uribe and Neil Hunt. 2015. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)* 6, 4 (2015), 1–19.

[68] Laura A Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 478–479.

[69] Bryan T Grenfell, Ottar N Bjørnstad, and Jens Kappey. 2001. Travelling waves and spatial hierarchies in measles epidemics. *Nature* 414, 6865 (2001), 716–723.

[70] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)* 53, 4 (2020), 1–37.

[71] Wenshuo Guo, Karl Krauth, Michael Jordan, and Nikhil Garg. 2021. The Stereotyping Problem in Collaboratively Filtered Recommender Systems. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–10.

[72] Kristina Halvorson and Melissa Rach. 2012. *Content Strategy for the Web: Content Strategy Web _p2*. New Riders.

[73] Mark S Handcock and Krista J Gile. 2010. Modeling social networks from sampled data. *The Annals of Applied Statistics* 4, 1 (2010), 5.

[74] Steve Hanneke, Wenjie Fu, and Eric P Xing. 2010. Discrete temporal models of social networks. *Electronic journal of statistics* 4 (2010), 585–605.

[75] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)* 5, 4 (2015), 1–19.

[76] Naieme Hazrati, Mehdi Elahi, and Francesco Ricci. 2020. Simulating the Impact of Recommender Systems on the Evolution of Collective Users' Choices. In *Proceedings ofthe 31st ACM Conference on Hypertext and Social Media (HT'20)*. 207–212. https://doi.org/10.1145/3372923.3404812

[77] Sherry He, Brett Hollenbeck, and Davide Proserpio. 2021. The market for fake reviews. *Available at SSRN 3664992* (2021).

[78] Dustin Hillard, Stefan Schroedl, Eren Manavoglu, Hema Raghavan, and Chirs Leggetter. 2010. Improving ad relevance in sponsored search. In *Proceedings of the third ACM international conference on Web search and data mining*. 361–370.

[79] Donna L Hoffman and Marek Fodor. 2010. Can you measure the ROI of your social media marketing? *MIT Sloan management review* 52, 1 (2010), 41.

[80] Russell Holly. 2012. The Play Store. In *Taking Your Android Tablets to the Max*. Springer, 71–88.

[81] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. 2018. Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 368–377.

[82] Dietmar Jannach and Christine Bauer. 2020. Escaping the McNamara Fallacy: Toward More Impactful Recommender Systems Research. *AI Magazine* 41, 4 (2020), 79–95. https://doi.org/10.1609/aimag.v41i4.5312

[83] Bernard J Jansen and Paulo R Molina. 2006. The effectiveness of Web search engines for retrieving relevant ecommerce links. *Information Processing & Management* 42, 4 (2006), 1075–1098.

[84] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[85] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 243–250.

[86] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2017. Accurately interpreting clickthrough data as implicit feedback. In *ACM SIGIR Forum*, Vol. 51. Acm New York, NY, USA, 4–11.

[87] Thorsten Joachims, Adith Swaminathan, and Tobias Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 781–789.

[88] Kwon Jung, Yoon C Cho, and Sun Lee. 2014. Online shoppers' response to price comparison sites. *Journal of Business Research* 67, 10 (2014), 2079–2087.

[89] Roger A Kerin, P Rajan Varadarajan, and Robert A Peterson. 1992. First-mover advantage: A synthesis, conceptual framework, and research propositions. *Journal of marketing* 56, 4 (1992), 33–52.

[90] Bernard Koch, Alex Hanna, Emily Denton, and Jacob Foster. 2021. Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.

[91] Karl Krauth, Sarah Dean, Alex Zhao, Wenshuo Guo, Mihaela Curmei, Benjamin Recht, and Michael I Jordan. 2020. Do Offline Metrics Predict Online Performance in Recommender Systems? *arXiv preprint arXiv:2011.07931* (2020).

[92] Haris Krijestorac, Rajiv Garg, and Vijay Mahajan. 2020. Cross-platform spillover effects in consumption of viral content: A quasi-experimental analysis using synthetic controls. *Information Systems Research* 31, 2 (2020), 449–472.

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

[94] Shiyang Lai and Ningyuan Fan. 2021. Understanding the attenuation of the accommodation recommendation spillover effect in view of spatial distance. *Journal of the Association for Information Science and Technology* (2021).

[95] Shyong K Lam and John Riedl. 2004. Shilling recommender systems for fun and profit. In *Proceedings of the 13th international conference on World Wide Web*. 393–402.

[96] Julian Lamont and Christi Favor. 2017. Distributive Justice. In *The Stanford Encyclopedia of Philosophy* (winter 2017 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[97] Bo Li, Yining Wang, Aarti Singh, and Yevgeniy Vorobeychik. 2016. Data poisoning attacks on factorization-based collaborative filtering. *arXiv preprint arXiv:1608.08182* (2016).

[98] Chen Liang, Zhan Shi, and TS Raghu. 2019. The spillover of spotlight: Platform recommendation in the mobile app market. *Information Systems Research* 30, 4 (2019), 1296–1318.

[99] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*. PMLR, 3150–3158.

[100] Lydia T Liu, Nikhil Garg, and Christian Borgs. 2021. Strategic ranking. *arXiv preprint arXiv:2109.08240* (2021).

[101] Tie-Yan Liu. 2011. Learning to rank for information retrieval. (2011).

[102] Zhi Liu and Nikhil Garg. 2021. Test-Optional Policies: Overcoming Strategic Behavior and Informational Gaps. , Article 11 (2021), 13 pages. https://doi.org/10.1145/3465416.3483293

[103] Zeyang Liu, Yiqun Liu, Ke Zhou, Min Zhang, and Shaoping Ma. 2015. Influence of vertical result in web search examination. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 193–202.

[104] Michael Luca. 2016. Reviews, reputation, and revenue: The case of Yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper* 12-016 (2016).

[105] Michael Luca and Georgios Zervas. 2016. Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science* 62, 12 (2016), 3412–3427.

[106] Eli Lucherini, Matthew Sun, Amy Winecoff, and Arvind Narayanan. 2021. T-RECS: A simulation tool to study the societal impact of recommender systems. *arXiv preprint arXiv:2107.08959* (2021).

[107] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback Loop and Bias Amplification in Recommender Systems. In *CIKM'20: International Conference on Information and Knowledge Management*. 2145–2148. https://doi.org/10.1145/3340531.3412152 arXiv:2007.13019

[108] Brian McFee, Thierry Bertin-Mahieux, Daniel PW Ellis, and Gert RG Lanckriet. 2012. The million song dataset challenge. In *Proceedings of the 21st International Conference on World Wide Web*. 909–916.

[109] James McInerney, Ehtsham Elahi, Justin Basilico, Yves Raimond, and Tony Jebara. 2021. Accordion: A trainable simulator forlong-term interactive systems. In *RecSys 2021 - 15th ACM Conference on Recommender Systems*. 102–113. https://doi.org/10.1145/3460231.3474259

[110] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[111] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a fair marketplace: Counterfactual evaluation of the trade-off between relevance, fairness & satisfaction in recommendation systems. In *Proceedings of the 27th acm international conference on information and knowledge management*. 2243–2251.

[112] Ali Mesbah, Arie Van Deursen, and Stefan Lenselink. 2012. Crawling Ajax-based web applications through dynamic analysis of user interface state changes. *ACM Transactions on the Web (TWEB)* 6, 1 (2012), 1–30.

[113] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *FAccT 2021 - Proceedings of the 2021 Conference on Fairness, Accountability, and Transparency*. 735–746.

[114] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting computer vision datasets: An invitation to reflexive data practices. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021), 161–172. https://doi.org/10.1145/3442188.3445880

[115] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8 (2021), 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

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

[117] Martin Mladenov, Chih-Wei Hsu, Vihan Jain, Eugene Ie, Christopher Colby, Nicolas Mayoraz, Hubert Pham, Dustin Tran, Ivan Vendrov, and Craig Boutilier. 2021. RecSim NG: Toward Principled Uncertainty Modeling for Recommender Ecosystems. (2021). arXiv:2103.08057 http://arxiv.org/abs/2103.08057

[118] Wendy W Moe and Peter S Fader. 2004. Dynamic conversion behavior at e-commerce sites. *Management Science* 50, 3 (2004), 326–335.

[119] Arash Molavi Kakhki, Chloe Kliman-Silver, and Alan Mislove. 2013. Iolaus: Securing online content rating systems. In *Proceedings of the 22nd international conference on World Wide Web*. 919–930.

[120] Faidra Georgia Monachou and Itai Ashlagi. 2019. Discrimination in online markets: Effects of social bias on learning from reviews and policy design. *Advances in Neural Information Processing Systems* 32 (2019), 2145–2155.

[121] Aadi Swadipto Mondal, Rakesh Bal, Sayan Sinha, and Gourab K Patro. 2021. Two-Sided Fairness in Non-Personalised Recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 15851–15852.

[122] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling fairness and bias in dynamic learning-to-rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 429–438.

[123] Tien T. Nguyen, Pik-Mai Hui, F. Maxwell Harper, Loren Terveen, and Joseph A. Konstan. 2014. Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity. In *Proceedings of the 23rd international conference on World wide web - WWW '14*. 677–686.

[124] NIST. 2004. TREC Data. https://trec.nist.gov/data.html. [Online; accessed May-2022].

[125] Giorgio Maria Di Nunzio, Alessandro Fabris, Gianmaria Silvello, and Gian Antonio Susto. 2021. Incentives for Item Duplication Under Fair Ranking Policies. In *International Workshop on Algorithmic Bias in Search and Recommendation*. Springer, 64–77.

[126] European Commission Independent High-Level Expert Group on Artificial Intelligence. 2019. *Ethics Guidelines for Trustworthy AI*. Retrieved August 10, 2021 from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

[127] Harrie Oosterhuis and Maarten de Rijke. 2018. Ranking for relevance and display preferences in complex presentation layouts. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 845–854.

[128] Gourab K Patro, Arpita Biswas, Niloy Ganguly, Krishna P Gummadi, and Abhijnan Chakraborty. 2020. Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In *Proceedings of The Web Conference 2020*. 1194–1204.

[129] Gourab K Patro, Abhijnan Chakraborty, Ashmi Banerjee, and Niloy Ganguly. 2020. Towards safety and sustainability: Designing local recommendations for post-pandemic world. In *Fourteenth ACM Conference on Recommender Systems*. 358–367.

[130] Gourab K Patro, Abhijnan Chakraborty, Niloy Ganguly, and Krishna Gummadi. 2020. Incremental fairness in two-sided market platforms: on smoothly updating recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 181–188.

[131] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web*. Springer, 325–341.

[132] Edward L Platt, Rahul Bhargava, and Ethan Zuckerman. 2015. The international affiliation network of YouTube trends. In *Ninth International AAAI Conference on Web and Social Media*.

[133] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 157–164.

[134] Kira Radinsky, Krysta M Svore, Susan T Dumais, Milad Shokouhi, Jaime Teevan, Alex Bocharov, and Eric Horvitz. 2013. Behavioral dynamics on the web: Learning, modeling, and prediction. *ACM Transactions on Information Systems (TOIS)* 31, 3 (2013), 1–37.

[135] Nimrod Raifer, Fiana Raiber, Moshe Tennenholtz, and Oren Kurland. 2017. Information retrieval meets game theory: The ranking competition between documents' authors. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 465–474.

[136] Manav Raj. 2021. Friends in High Places: Demand Spillovers and Competition on Digital Platforms. *Available at SSRN 3843249* (2021).

[137] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithmic impact assessments: A practical framework for public agency accountability. *AI Now Institute* April (2018), 22. https://ainowinstitute.org/aiareport2018.pdf

[138] Rejoiner. 2018. The Amazon Recommendations Secret to Selling More Online. (2018).

[139] Marco Tulio Ribeiro, Anisio Lacerda, Edleno Silva, D E Moura, Edleno Silva De Moura, Itamar Hata, Adriano Veloso, and Nivio Zi. 2013. Multi-Objective Pareto-Efficient Approaches for Recommender Systems. *ACM Trans. Intell. Syst. Technol* 9, 1 (2013). https://doi.org/10.1145/2629350

[140] Stephen E Robertson. 1977. The probability ranking principle in IR. *Journal of documentation* (1977).

[141] David Rohde, Stephen Bonner, Travis Dunlop, Flavian Vasile, and Alexandros Karatzoglou. 2018. RecoGym: A Reinforcement Learning Environment for the problem of Product Recommendation in Online Advertising. (2018). arXiv:1808.00720 http://arxiv.org/abs/1808.00720

[142] Matthew J Salganik and Duncan J Watts. 2008. Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social psychology quarterly* 71, 4 (2008), 338–355.

[143] Marlesson R. O. Santana, Luckeciano C. Melo, Fernando H. F. Camargo, Bruno Brandão, Anderson Soares, Renan M. Oliveira, and Sandor Caetano. 2020. MARS-Gym: A Gym framework to model, train, and evaluate Recommender Systems for Marketplaces. In *2020 International Conference on Data Mining Workshops (ICDMW)*. 189–197. https://doi.org/10.1109/ICDMW51313.2020.00035

[144] Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists. In *Companion Proceedings of The 2019 World Wide Web Conference*. 553–562.

[145] Murugesan Saravanakumar and T SuganthaLakshmi. 2012. Social media marketing. *Life science journal* 9, 4 (2012), 4444–4451.

[146] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37. https://doi.org/10.1145/3476058 arXiv:2108.04308

[147] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*. PMLR, 1670–1679.

[148] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2018. Fairness and Abstraction in Sociotechnical Systems. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, Vol. 1. 59–68. https://doi.org/10.1145/3287560.3287598

[149] Amit Sharma, Jake M Hofman, and Duncan J Watts. 2015. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*. 453–470.

[150] Jing-Cheng Shi, Yang Yu, Qing Da, Shi-Yong Chen, and An-Xiang Zeng. 2019. Virtual-taobao: Virtualizing real-world online retail environment for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 4902–4909.

[151] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2219–2228.

[152] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2239–2248.

[153] Harald Steck. 2011. Item popularity and recommendation accuracy. In *Proceedings of the fifth ACM conference on Recommender systems*. 125–132.

[154] Tom Sühr, Asia J Biega, Meike Zehlike, Krishna P Gummadi, and Abhijnan Chakraborty. 2019. Two-sided fairness for repeated matchings in two-sided markets: A case study of a ride-hailing platform. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3082–3092.

[155] Tom Sühr, Sophie Hilgard, and Himabindu Lakkaraju. 2020. Does Fair Ranking Improve Minority Outcomes? Understanding the Interplay of Human and Algorithmic Biases in Online Hiring. *arXiv preprint arXiv:2012.00423* (2020).

[156] Özge Sürer, Robin Burke, and Edward C Malthouse. 2018. Multistakeholder recommendation with provider constraints. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 54–62.

[157] Moshe Tennenholtz and Oren Kurland. 2019. Rethinking search engines and recommendation systems: a game theoretic perspective. *Commun. ACM* 62, 12

(2019), 66–75.
[158] TrustPilot. 2020. *Action We Take.* Retrieved January 10, 2022 from https://uk.legal.trustpilot.com/for-everyone/action-we-take
[159] European Union. 2016. Data minimization principle. https://gdpr-info.eu/art-5-gdpr/. [Article 5(1)(c) of the GDPR and Article 4(1)(c) of Regulation (EU) 2018/1725].
[160] Briana Vecchione, Solon Barocas, and Karen Levy. 2021. Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21)*, Vol. 1. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3465416.3483294 arXiv:2109.06974
[161] Patrick Vonderau. 2019. The Spotify effect: Digital distribution and financial growth. *Television & New Media* 20, 1 (2019), 3–19.
[162] Ellen M Voorhees. 1999. The TREC-8 question answering track report. In *Trec*, Vol. 99. Citeseer, 77–82.
[163] Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management* 36, 5 (2000), 697–716.
[164] Xuanhui Wang, Nadav Golbandi, Michael Bendersky, Donald Metzler, and Marc Najork. 2018. Position bias estimation for unbiased learning to rank in personal search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 610–618.
[165] Yixin Wang, Dawen Liang, Laurent Charlin, and David M Blei. 2020. Causal inference for recommender systems. In *Fourteenth ACM Conference on Recommender Systems*. 426–431.
[166] Yue Wang, Dawei Yin, Luo Jie, Pengyuan Wang, Makoto Yamada, Yi Chang, and Qiaozhu Mei. 2016. Beyond ranking: Optimizing whole-page presentation. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*. 103–112.
[167] Lin Xiao, Zhang Min, Zhang Yongfeng, Gu Zhaoquan, Liu Yiqun, and Ma Shaoping. 2017. Fairness-aware group recommendation with pareto-efficiency. In *Proceedings of the eleventh ACM conference on recommender systems*. 107–115.
[168] Leyang Xue, Peng Zhang, and An Zeng. 2019. Enhancing the long-term performance of recommender system. *Physica A: Statistical Mechanics and its Applications* 531 (2019), 121731.
[169] Tao Yang and Qingyao Ai. 2021. Maximizing Marginal Fairness for Dynamic Learning to Rank. In *Proceedings of the Web Conference 2021*. 137–145.
[170] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A Survey on Causal Inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 5 (2021), 1–46.
[171] Sirui Yao, Yoni Halpern, Nithum Thain, Xuezhi Wang, Kang Lee, Flavien Prost, Ed H. Chi, Jilin Chen, and Alex Beutel. 2020. Measuring Recommender System Effects with Simulated Users. In *Second Workshop on Fairness, Accountability, Transparency, Ethics and Society on the Web.* arXiv:2101.04526 http://arxiv.org/abs/2101.04526
[172] Sirui Yao and Bert Huang. 2017. Beyond Parity: Fairness Objectives for Collaborative Filtering. *Advances in Neural Information Processing Systems* 30 (2017), 2921–2930.
[173] YouTube. 2018. *Preventing Harm to the Broader YouTube Community.* Retrieved January 10, 2022 from https://blog.youtube/news-and-events/preventing-harm-to-broader-youtube/
[174] Quan Yuan, Gao Cong, Zongyang Ma, Aixin Sun, and Nadia Magnenat Thalmann. 2013. Time-aware point-of-interest recommendation. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 363–372.
[175] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th international conference on World wide web*. 1011–1018.
[176] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.
[177] Meike Zehlike and Carlos Castillo. 2020. Reducing disparate exposure in ranking: A learning to rank approach. In *Proceedings of The Web Conference 2020*. 2849–2855.
[178] Meike Zehlike, Tom Sühr, Ricardo Baeza-Yates, Francesco Bonchi, Carlos Castillo, and Sara Hajian. 2022. Fair Top-k Ranking with multiple protected groups. *Information Processing & Management* 59, 1 (2022), 102707.
[179] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2021. Fairness in Ranking: A Survey. (2021), 39–4. arXiv:arXiv:2103.14000v1
[180] Hengtong Zhang, Yaliang Li, Bolin Ding, and Jing Gao. 2020. Practical Data Poisoning Attack against Next-Item Recommendation. In *Proceedings of The Web Conference 2020*. 2458–2464.
[181] Jingjing Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang Ketter. 2020. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research* 31, 1 (2020), 76–101. https://doi.org/10.1287/ISRE.2019.0876
[182] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. 2020. Long-term impacts of fair machine learning. *Ergonomics in Design* 28, 3 (2020), 7–11.