Toward A Two-Sided Fairness Framework in Search and Recommendation

Jiqun Liu
jiqunliu@ou.edu
The University of Oklahoma
Norman, OK, USA

ABSTRACT

As artificial intelligence (AI) assisted search and recommender systems have become ubiquitous in workplaces and everyday lives, understanding and accounting for *fairness* has gained increasing attention in the design and evaluation of such systems. While there is a growing body of computing research on measuring *system fairness and biases* associated with data and algorithms, the impact of *human biases* that go beyond traditional machine learning (ML) pipelines still remain understudied. In this Perspective Paper, we seek to develop a *two-sided* fairness framework that not only characterizes *data and algorithmic biases*, but also highlights the *cognitive and perceptual biases* that may exacerbate system biases and lead to unfair decisions. Within the framework, we also analyze the interactions between human and system biases in search and recommendation episodes. Built upon the two-sided framework, our research synthesizes intervention and intelligent nudging strategies applied in cognitive and algorithmic debiasing, and also proposes novel goals and measures for evaluating the performance of systems in addressing and proactively mitigating the risks associated with biases in data, algorithms, and bounded rationality. This paper uniquely integrates the insights regarding human biases and system biases into a cohesive framework and extends the concept of fairness from human-centered perspective. The extended fairness framework better reflects the challenges and opportunities in users' interactions with search and recommender systems of varying modalities. Adopting the two-sided approach in information system design has the potential to enhancing both the *effectiveness* in online debiasing and the *usefulness* to boundedly rational users engaging in information-intensive decision-making.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**.

KEYWORDS

Two-sided fairness, human bias, system bias, information retrieval, recommender system

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHIIR '23, March 19–23, 2023, Austin, TX, USA
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0035-4/23/03…$15.00
https://doi.org/10.1145/3576840.3578332

ACM Reference Format:
Jiqun Liu. 2023. Toward A Two-Sided Fairness Framework in Search and Recommendation. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3576840.3578332

1 INTRODUCTION

Artificial Intelligence (AI) assisted search and recommender systems have become ubiquitous in workplaces and everyday lives, and play a significant role in human decision-making activities. However, the underlying algorithms and data could be *unfair* and skewed toward a particular community or group of people, leading to biased judgments and problematic decisions. For instance, COMPAS, a software used by the courts in the United States to estimate the risk of a person to recommit another crime, is more likely to have higher false positive rates in predicting the recidivism of African-American offenders [1]. Also, AI systems built upon medical and usage data mainly collected from men could falsely underestimate the risk of heart attack faced by women, which aggravates gender inequality in health [2]. The AI-assisted retrieval algorithms (e.g. BERT [12, 22]) behind Web search engines face similar problems as they could be picking up on biases from data providers, algorithm designers, and users "in the way a child mimics the bad behavior of his parents" [3]. Given these sociotechnical challenges, a growing body of computing research strives to measure *system-side fairness* and mitigate the risks of biases embedded in algorithms and training data [56]. These increasing research efforts give rise to a series of relevant workshops, grants, and emerging communities (e.g. ACM FAccT [4]).

1.1 Biased Systems and Boundedly Rational Users

While existing research has achieved significant progresses in measuring and mitigating system bias in a broad range of application scenarios, the impact of *human bias* that goes beyond traditional machine learning (ML) pipelines still remains understudied. According to Kahneman [39], human bias refers to the *systematic deviations* of human behavior from the predictions of rational normative models. In contrast to the assumptions of many simulated user models, people are *boundedly rational* and their decisions are often affected by a series of biases and mental shortcuts [72]. Thus, when interacting

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[2] https://www.forbes.com/sites/carmenniethammer/2020/03/02/ai-bias-could-put-womens-lives-at-riska-challenge-for-regulators
[3] https://www.nytimes.com/2019/11/11/technology/artificial-intelligence-bias.html
[4] https://facctconference.org/.

with search and recommender systems, a user may be disproportionately impacted by the retrieved results or recommended items that trigger the behavioral impact of their existing cognitive biases and heuristics, causing unexpected unfair outcomes. For example, users who have certain misleading beliefs regarding vaccines are more vulnerable to related misinformation presented on search engine result pages (SERPs), leading to ill-informed, undesired medical decisions. Online shoppers tend to quickly accept immediate mediocre recommendations after encountering several bad-quality products recommended by the system (as low reference levels). Differing from data and algorithmic biases, human biases tend to be individualized, context-dependent [40, 75], and closely associated with people's previous similar experiences (e.g. case-based decision making [29]). However, both system bias and human bias could result in unfair decisions and negative societal impacts. How to identify and mitigate the risks of potential biases from both sides is a fundamental open challenge to information retrieval (IR) and recommender systems (RS) communities.

1.2 Two-sided Fairness Perspective

To address the gap above, we re-conceptualize fairness in AI from a *user-centered* perspective and propose a *two-sided fairness framework* that deconstructs the impact of both system bias and human bias in interactive search, recommendation, and AI-assisted decision making. Aligned with the objectives of CHIIR Perspective Paper Track, our work seeks to present novel insights and identify open questions at conceptual, methodological, and evaluation levels:

- We **extend the concept of fairness** to cover the effects and measurements of both human bias and system bias embedded in data and algorithms, as well as the possible interactions between them.
- We **propose new two-sided evaluation methods** that can examine the performance of search and recommender systems in addressing and proactively reducing the impacts of both human and system biases.
- We **synthesize empirically tested re-ranking, intervention and nudging techniques** that could potentially mitigate the risks of one or both types of biases.

Through accomplishing the above goals, this paper makes threefold contributions: *(i)* It integrates the interdisciplinary insights from IR and recommendation, AI fairness, and cognitive psychology and offers a more balanced, psychologically realistic approach to measuring and evaluating fairness in users' interactions with intelligent information systems; *(ii)* It highlights the available tools (e.g. recommendation and re-ranking, system intervention, intelligent nudging techniques) for mitigating the risks of system and human biases in information-intensive tasks; *(iii)* It proposes novel evaluation metrics that measure the performance of systems in reducing both system and human biases and could be employed and tested in a broader scope of search and recommendation scenarios. In addition, this paper also identifies new fundamental and empirical issues that emerge from the extended fairness concept and have the potential to inspire substantive discussions and significant progresses in the field.

2 EXTENDING THE SCOPE OF FAIRNESS

Going beyond the mainstream studies on algorithmic fairness, this section presents an extended definition of fairness that sets constraints on both system bias and human bias in users' interactions with systems. Under the extended concept, the fairness of a system will be evaluated based on not only its performance in *reducing the biases inherited from data and algorithms*, but also its ability in *protecting users from the risks and contextual triggers of cognitive and perceptual biases*. Our two-sided fairness framework incorporates the features of biases from both sides and speaks to new challenges in understanding and supporting boundedly rational users interacting with potentially biased systems.

2.1 Notions of System Fairness

Search and recommender systems have been employed by a growing user population as the main channel for information access in varying tasks, including the ones in sensitive environments, such as health information seeking [1], hiring and job application [63, 80], and financial decision making [7]. Thus, the underlying biased data and unfair algorithm would not only affect information presentation, but also lead to unfair distributions of economic and socio-technical resources.

To address this, inspired by classic research on fairness from psychology and philosophy [e.g. 14, 37], researchers proposed a series of fairness definitions focusing on varying levels and factors, and employed them as constraints to mitigate the bias and discrimination in retrieval and recommendation algorithms. According to [56], the existing fairness definitions can be grouped into three categories: Individual fairness, group fairness, and subgroup fairness. *Individual Fairness* requires that systems should give similar predictions to individual users and content generators with similar characteristics, regardless of their differences in protected sensitive attributes, such as gender, ethnicity, and popularity [13, 23, 45]. *Group Fairness* concept focuses on the potential biases against sensitive groups or communities and emphasizes that all groups should be treated equally [23, 24, 45]. *Subgroup Fairness* combines the features of both fairness concepts above and measures whether a fairness constraint holds over a large set of subgroups [41, 42]. Existing fairness concepts and measures seek to mitigate and prevent varying types of *observable unfairness* in retrieved contents and recommended items, especially with respect to certain protected attributes. However, the potential risks of *implicit unfairness* generated through the combination of system output and user biases still remain unclear.

2.2 Human Bias and Bounded Rationality

Differing from (over)simplified simulated agents seeking maximized utility, real-life users often operates under the impact of cognitive and perceptual biases, and attempt to *satisfice* or achieve *good-enough* results, rather than optimize [6, 39, 72]. Human biases and satisficing strategies, covered under the theoretical umbrella of *bounded rationality*, could drive users to unconsciously make biased judgments regarding retrieved information and recommended items and unfair decisions in sensitive environments. Current fairness metrics and constraints focusing on the biases in data, retrieval algorithms, and recommendation mechanisms are widely applied

in standardized offline experiments and are independent from user characteristics by design. Thus, extending existing fairness concepts to cover human biases would be essential, especially for the scenarios where algorithmically fair systems still result in *practically unfair* information use and decisions. For instance, users may click and save search results that are consistent with their pre-search expectations and opinions of certain cognitive authorities, despite the diverse set of topics, perspectives and content generators included on the SERP (Confirmation Bias [70]). In addition, diverse recommended items ranked on similar positions may receive significantly different amounts of actual attention due to the divergence in users' remembered experiences with similar products (Reference Dependence [51]). Given this challenge, it is critical to enrich the fairness concept with user dimensions and evaluate the performance of search and recommender systems in identifying, preventing, and mitigating the negative effect of human bias. At methodological level, adopting certain intervention and nudging techniques for cognitive debiasing may not only address immediate biased judgments in current interactions, but also generate sustained impacts on future information searching and recommendation assessments [52, 53, 87].

2.3 Interaction between Human Bias and System Bias

Human bias and system bias can interact with each other at different stages of users' interactions with IR and recommender systems [6], such as search initiation and query reformulation, browsing and clicking, and evaluation of information items and products. For instance, Ge et al. [28] studied the interactions between users' interests as *anchoring level* and personalized e-commerce recommendations. In particular, based on the interaction logs gathered from Alibaba Taobao transactions, researchers measured the self-reinforcement effect on users' interests caused by the narrowed exposure of recommended product types. This mutual reinforcement between users' initial preferences and the customized recommendations tailored according to in-situ behaviors could lead to *echo chamber* effect. In IR evaluation, Scholer et al. [69] investigated the dynamic thresholds in external assessors' document judgments and their associations with the sequence of presenting documents of varying relevance. The results indicate that initially encountered high-quality documents may heighten a user's reference level of relevance, leading to underestimated relevance levels in subsequent document judgments. In addition to human-system interactions, Azzopardi [6] also argues that information searchers may experience mixed effects of multiple cognitive biases in search and evaluation.

When interacting with the same set of algorithmically fair results, different users may have significantly different chances of making biased judgments due to their differences in pre-interaction references and expectations, remembered experiences, and in-situ perceived gains and efforts. Users' biases and system biases may reinforce each other through implicit feedback, learning to rank (LTR), and personalized recommendation processes. While human biases are difficult to observe and often act unconsciously, they could still cause unfair decisions and tangible consequences for people with different beliefs, knowledge bases, and prior experiences. To

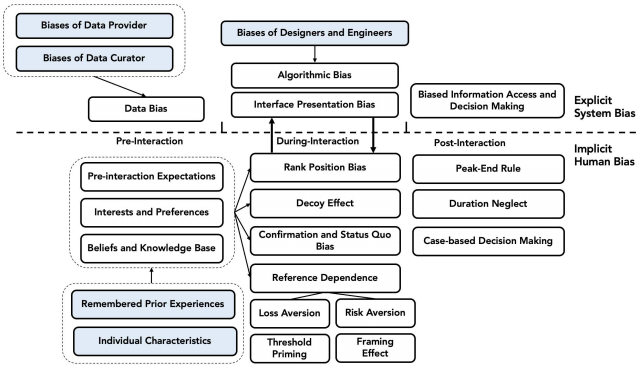

Figure 1: Two-sided bias structure in users' interactions with search and recommender systems.

make matter worse, users' biases are often purposely exploited for increasing engagements and profits, especially in online shopping, social media feeds, and marketing promotions [33, 77, 86], leading to unseen unfairness. Compared to widely discussed protected attributes [cf. 18], factors associated with human biases are usually hidden in fast judgments and intuitive decisions, and are closely related to local contexts (e.g. search intention, domain knowledge, cognitive load) and individual characteristics (e.g. short-term memory span). Thus, understanding and achieving *human-centered fairness* would be more empirically challenging but also equally important to reaching system fairness (especially the AI/Machine Learning components [e.g. 5, 19]) in an era of information ubiquity.

3 A TWO-SIDED FAIRNESS FRAMEWORK

Built upon above discussions, this section proposes a *two-sided fairness framework* that takes into consideration the features, effects, and measurements of both human bias and system bias. The extended fairness concept can inspire and inform a more balanced, user-aware approach to evaluating the fairness of search and recommender systems.

Figure 1 illustrates the biases from both human and system sides that may operate at different stages of user interactions. Given that the mainstream fairness research focusing on algorithmic fairness in AI/ML, our framework presents a balanced approach to addressing two-sided biases at different stages (i.e. pre-interaction, interaction, post-interaction), with an emphasis on the impacts of human biases. Note that our work discusses the major types of human and system biases, especially the ones that are empirically examined in search and recommendation contexts [e.g. 6], and is not able to exhaust all possible biases. A more comprehensive list of human bias is offered by Benson [5].

3.1 Pre-Interaction Stage

Many of the conditions and triggers of human bias and system bias are formed long before users' interactions with systems actually occur. On the system side, biases in the data employed in training

[5]Cognitive bias cheat sheet: https://betterhumans.pub/cognitive-bias-cheat-sheet-55a472476b18.

ML-based algorithms (e.g. LTR, adaptive recommendation) may result in biased algorithmic outcomes. The biases in training data may originate from the biased sampling and curation processes (which creates non-representative samples), as well as the existing historical bias and socio-technical problems in reality [74]. Data bias and unfairness may also occur due to human prejudice and stereotyping based upon sensitive attributes [56]. In addition, part of the data bias could result from biased behaviors during interactions across varying recommendation platforms and search interfaces [61]. For instance, users may spend more time and clicks on the results and items that are ranked on the top of SERPs or consistent with their expectations, which could generate skewed feedback data and reinforce existing biases in relevance and usefulness estimations [2, 55].

On the human side, Users' pre-search expectations, interests and preferences, as well as beliefs and knowledge base are affected by their remembered prior experiences under similar scenarios (or "cases" in CBDT [29]) and individual characteristics. For example, users may choose to avoid certain information sources or vendors due to previous negative experience under similar scenarios. Also. users who lack certain domain knowledge may skip unfamiliar or seemingly ambiguous results on SERPs [60]. These cognitive factors usually shape the reference levels based on which users evaluate available options during interactions, such as retrieved results, recommended queries and products, as well as search continuation or stopping (*Reference Dependence* bias [79]). With divergent reference levels, users interacting with similar result lists tend to perceive and evaluate information gains and search efforts differently, resulting in distinct search tactics, judgments, and decision-making strategies [15, 51]. In addition, the pre-interaction factors could also affect the extent to which a user is vulnerable to the negative impact of other potential biases during interactions. For instance, a medical expert may be less likely to be influenced by the vaccine misinformation ranked on top positions of SERPs compared to novice searchers in the field. A computer scientist who is familar with personalized recommendation algorithms may possess a high level of *algorithm awareness* [cf. 34] and could be more sensitive to biased recommendations from (over)personalized systems. Investigating pre-interaction factors and associated human biases will allow researchers to better understand why users with different backgrounds and prior interaction experiences may have significantly different likelihood of achieving optimal utility or desired outcomes when facing similar sets of information and recommendations.

3.2 Interaction Stage

On the system side, a biased algorithmic decision (e.g. music recommendations that do not provide fair representation of new artists; global economy search results that mainly focuses on a small set of developed economies) could happen because of both data biases inherited through training and built-in biases embedded in algorithms [56]. In existing research on system and algorithmic fairness, fairness is often broadly defined as the "absence of prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making" [67]. According to [35], this general rule of fairness can be written as:

$$P(\hat{Y} = 1|M = 0, Y = y) = P(\hat{Y} = 1|M = 1, Y = y), y \in \{0, 1\} \quad (1)$$

Where $\hat{Y}$ represents the predicted results, and M as a binary variable indicates if the data point represents a protected group member. The goal of this *Equalized-Odds* fairness constraint is that the probability of an item in the positive category being correctly assigned a positive label and the probability of an item from the negative class being incorrectly put into positive category should stay the same regardless of the protected membership labels [56, 81]. For instance, with the same level of actual relevance and quality, the contents and products produced by both popular and new providers should obtain equalized likelihood of exposure on similar rank positions. Also, when searching under a controversial topic, users should have access to fairly distributed information sources with diverse perspectives.

The *Equalized-Odds* fairness measure has also been adjusted according to specific application scenarios and fairness requirements. For example, *Equalized-Opportunity* fairness focuses on the positive labels and requires that the probability of an item in a positive class being labeled with a positive outcome should be equal for both protected and unprotected group members [35]. This constraint can be written as:

$$P(\hat{Y} = 1|M = 0, Y = 1) = P(\hat{Y} = 1|M = 1, Y = 1) \quad (2)$$

Similar rules can also be applied to group and subgroup fairness research, where equal true positive and false positive rates should be achieved in ML-based predictions for groups with different protected attribute labels. Mitigating and restricting behind-the-scenes algorithmic biases could result in a fairer presentation of information and recommendations accessed by users. Apart from the abstracted biases in predictions, researchers have also explored potential system biases in interface design and information presentation [60, 68] that could trigger some users' misapplication of mental shortcuts and heuristics [39, 75] and thereby cause obstacles in inferential thinking and information evaluation.

Differing from data and algorithmic biases for which external labeling is relatively straightforward, human biases that emerge from and operate in interactions are often difficult to measure. While researchers could design diverse experimental conditions as assigned triggers of cognitive biases in controlled lab settings [39, 85], it is challenging to measure biases in users' real-time interactions with search and recommender systems under ill-defined, complex tasks. However, it is critical to study human bias and incite discussions on human-side fairness as some users may end up in implicitly disadvantaged positions in their interactions with systems due to their cognitive and perceptual biases being triggered by certain contextual factors, system outputs, and individual traits.

Figure 2 presents a hypothetical example of search/recommendation iteration under query or question i to illustrate the interrelated human biases that operate during interaction sessions. For instance, when interacting with retrieved products and information items, users may prefer to examine and click the ones that confirm their pre-interaction expectations and beliefs or are less likely to challenge their existing *status quo* in mind (*Confirmation and Status Quo Bias* [66]), which could reduce the probability of *cognitive dissonance* [4, 47] or knowledge restructuring. In addition, the initially

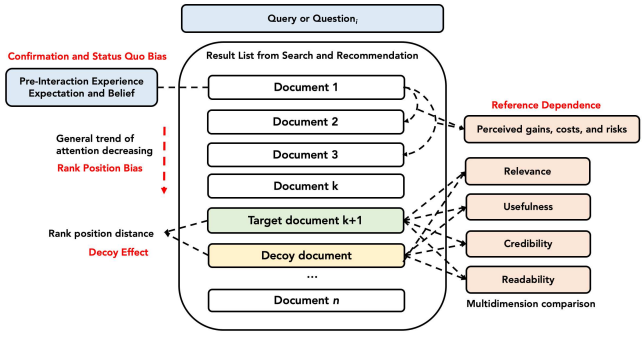

Figure 2: Interrelated human biases within interactions.

encountered results and recommended items could be used as reference or anchoring points in users' evaluation of following items, especially in terms of the perceived gains and costs involved in browsing and examination (*Reference Dependence*). Thus, when a user encounters a high-quality item at the beginning of the iteration (e.g. a highly relevant document or five-start product with good price), the user may form a relatively high reference level or in-situ expectation in mind. As a result, a slight drop in item quality or small increases in efforts (e.g. dwell time, number of clicks, and recommendations examined) may lead to a major decrease of interaction satisfaction in following browsing and clicking activities (*Loss Aversion Bias and Threshold Priming* [51, 69, 79]). Also, users may choose to avoid the search results that are framed as an ambiguous or unfamiliar, risky option (*Risk Aversion and Framing Effect* [43, 54, 62]). However, when the user starts with a low reference level or expectation, their perceptions and evaluations of subsequent items may change completely despite that the nature of the items stay the same.

Decoy Effect refers to the scenarios where people change their preference between two existing options when presented with a third option (i.e. the decoy) that is asymmetrically dominated [78, 88]. In IR and Crowdsourcing labeling, Eickhoff [25] examined the impact of decoy document on users' thresholds and strategies in standard relevance judgments. For instance, in a convenient store, a customer may find it difficult to decide between an apple and a banana for afternoon snack. However, when a rotten apple (the decoy) is placed next to the existing apple, the customer may find the apple to be a more favorable option as there is a perceived gain compared to the decoy reference. As Figure 2 shows, it may be difficult to predict a user's preference between document K and document $K+1$ as they are associated with two different subtopics, T_K and T_{K+1}, respectively. However, when a symmetrically dominated document D_{decoy} is presented and the topic is similar to T_{K+1}, it is likely that the user will give a higher score to the target document $K+1$. In IR, a user may compare the target and decoy results over multiple dimensions, such as relevance, usefulness, perceived credibility, and readability. The essence of decoy effect is that a person's preference between two or more options could be altered completely by adding a decoy option, without changing the nature of existing options. Similar impacts of decoy has also been empirically confirmed in online recommendation and e-Commerce settings [e.g. 71, 86].

Differing from the human biases introduced above, *Rank Position bias* [83] is easier to observe on the surface of interactions and has been discussed in a wide range of IR (particularly unbiased LTR [3]) and RS experiments [20, 32]. The knowledge regarding rank position bias has been widely applied in simulating user models underlying offline evaluation metrics, where users' attention and likelihood of clicking and examination are often assumed to be decreasing by rank in SERP and recommendation evaluations [e.g. 17, 58]. However, the actual effect of rank position bias may be moderated by the form and modality of search results and recommendations. For instance, researchers found that compared to organic search results, vertical results and recommendations (e.g. News, images) may appear to be more visually salient and reduce the impact of rank positions on the probability of examination and clicking [82].

As it is discussed above, human biases could be triggered by a series of pre-interaction factors and within-interaction factors (e.g. rank position, decoy items, initially encountered items, distance and similarity between results). Once triggered, human biases could lead to significant deviations of users' behaviors and judgments from optimal or desired results. Consequently, unfair decisions and outcomes may occur between users who are more vulnerable to certain biases and contextual triggers and the ones who are not. By extending existing fairness concepts, our two-sided fairness framework seek to highlight, characterize, and assess this human-side unfairness in search and recommender systems.

3.3 Post-Interaction Stage

At the post-interaction stage, the mixed effect of system biases and human biases may result in biased information evaluation and use, unfair decisions and undesired outcomes. On the system side, a biased algorithmic decision could come from a black-box re-ranking or recommendation model where the training data (generated at pre- and within-interaction stages) and learning algorithms could not be modified or scrutinized. On the human side, in addition to the within-interaction biases, users may subject to the influence of other *whole-session* cognitive biases when making decisions based on remembered experiences. For instance, when evaluating and comparing the performances of multiple queries and systems, users are heavily influenced by the *peak point* and *end or most-recent point* of experience during the sessions being evaluated and are not sensitive to the *actual time duration* of the interaction (*Peak-End Rule* and *Duration Neglect* [36, 51, 64]). These memory-related biases may lead to significant divergence between users' retrospective satisfaction-based judgments and the assessment from system designers and lead to unfair evaluations of systems and interaction experiences from whole-session contexts.

3.4 Two-Sided Fairness Goals

The system-side fairness goal can be adapted from current fairness objectives in AI/ML fairness, as it is indicated in sub-section 3.2. Despite the difference in specific measures, the common underlying goal is to achieve equal true positive and false positive prediction rates for both protected and unprotected group members identified with pre-defined protected attributes. The prediction results are

usually associated with critical decisions and societal impacts, such as healthcare, hiring, and house mortgage approval [48, 56].

Regarding human-side fairness, we can adopt similar approach and write it as:

$$P(Y = Y^*|M = 0, A = a) = P(Y = Y^*|M = 1, A = a) \quad (3)$$

Where Y^* represents the desired or accessible optimal outcome for an individual or group, A represents the set of attributes and contextual features that are not associated with human biases. M indicates if the user belongs to the protected group that is more vulnerable to certain cognitive and perceptual biases. Y represents the actual outcome or utility of information use and decision-making. The human-side fairness goal is that users with similar intentions of interacting with search and recommender systems and backgrounds should have similar chance of obtaining desired outcomes from the interaction, regardless of their actual vulnerability to the human biases that could be triggered. The group membership variable M can be written as:

$$M = M_{hb_1}, M_{hb_2}, M_{hb_3} ... M_{hb_m} \quad (4)$$

$$P(M_{hb_1} = 1) = f_M^{hb_1}(I_i, E_i, T_s) \quad (5)$$

Where M_{hb_m} represents a user's estimated vulnerability to a potential human bias hb_m. Without loss of generality, here we assume that every M_{hb} is a binary variable. The probability that a user is vulnerable to hb_1 is determined by the function, $f_M^{hb_1}$, of three variables: individual characteristics of the user i (I_i), prior experiences, expectations and beliefs E_i, and the triggers from system outputs T_s. As it is explained in previous sub-sections, different human biases may involve diverse mechanisms and probabilities of being triggered [39] and thus should be represented with separate functions. In addition, researchers and system designers should also explore the interplay of varying biases at different stages and investigate if addressing certain human biases would mitigate or increase the risk of encountering other biases in query reformulation, judgment of results and recommendations, as well as post-interaction decision-making.

3.5 Two-Sided Fairness in Human-Centered System Evaluation

Based on the goals defined above, researchers can evaluate the performance of search and recommender systems in fulfilling associated fairness constraints [e.g. 27, 81, 84]. Regarding system fairness, the predictions and output (e.g. ranked results, customized recommendations) of systems can be evaluated according to the measures specified in sub-section 3.2. For instance, when making algorithmic decisions on music recommendations, it is critical to assess if the probabilities and rank positions of recommending relevant musics are equal across artists from varying backgrounds [57]. In addition to this active approach, the system fairness constraint may also be achieved through *Unawareness*: an algorithm can be considered fair if no protected attribute is explicitly adopted in making decisions [31].

On the human side, the evaluation needs to be built upon a series of preparation work. Specifically, one have to complete following tasks before assessing human-centered fairness:

- **T1**: Predicting desired outcomes or estimating optimal outcomes based on the nature of tasks and problems that motivate users to engage with systems.
- **T2**: Estimating the real-time risk of cognitive and perceptual biases for individuals and groups based on the knowledge of user characteristics and features of system outputs.
- **T3**: Learning bias-aware user models to characterize users' information evaluation, use, and decision-making patterns under the impact of biases.

Among the above tasks, T1 will offer ground truth labels for evaluating the fairness in the probability of obtaining desired or optimal outcomes. In well-structured tasks with clearly-defined goals, the labels could also be extracted from users' annotations of their task goals. T2 will generate the protected attribute labels and classify individuals into multiple categories among which human-centered fairness needs to be achieved. Accomplishing T2 and developing bias-aware interventions will enhance equal access to quality information and facilitate unbiased judgments of encountered information, and thereby contribute to the completion of T1. Models built under T3 will allow us to predict potential biased judgments and decisions based on the estimated risks of biases, individual traits, and in-situ contextual triggers. With these models, researchers could proactively identify biased behaviors *before* problematic decisions actually occur. Under these human-centered fairness constraints, when evaluating SERPs and ranked list of recommended items, we should not only measure the explicit biases associated with protected sensitive attributes, but also estimate the risk of them in triggering varying types of human biases introduced in Figure 2. Depending on the specific systems a user interacts with and the nature of motivating tasks, the weights of human-side fairness and system-side fairness could be tailored to varying evaluation preferences.

3.6 Two-sided Fair Ranking, Intervention and Intelligent Nudging

In previous IR, ML, and Human-Computer Interaction (HCI) studies, a series of re-ranking, intervention and nudging techniques have been developed for system bias mitigation and cognitive debiasing, which could help enhance two-sided fairness. On the system and algorithmic side, most bias mitigation methods are developed in ML pipeline and can be grouped into three categories or stages: Pre-processing, in-processing, and post-processing [56]. *Pre-processing* category covers the techniques used in reducing and removing bias and discrimination in datasets employed for training ranking and recommendation algorithms [10]. For instance, researchers can apply preferential sampling methods to address discrimination in search and recommendation logs and ensure fair representation of samples from diverse communities and populations before training rankers. *In-processing* group includes the techniques for modifying learning algorithms and removing biases *during* interaction and model training processes [11]. For instance, IR researchers can adjust LTR algorithms (e.g. with counterfactual methods) and mitigate the possible biases and noise learned from historical data and user behaviors during search sessions (e.g. rank position bias) [3, 38].

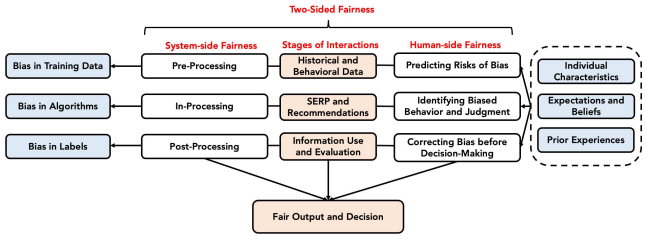

Figure 3: Enhancing Two-Sided Fairness.

Post-processing methods are applied after training and allow systems to reassign labels created by potentially biased black-box models [21].

On human-centered fairness side, a series of intelligent intervention and nudging techniques have been proposed and empirically tested in search, recommendation, and a diverse set of HCI scenarios. For instance, systems could transform digital information and add visual aids on interfaces to reduce the perceived ambiguity and increase saliency of certain information [16, 73]. Also, effective nudging and debiasing could be achieved through changing the decision structure, such as altering the starting and anchoring options, proactively adjust the ranking structure that could lead to negative decoy effects, and re-arranging evaluation sequences to prevent potential priming effects and reference dependence bias [16, 69, 76]. In addition, researchers could leverage the power of cognitive authority and query priming techniques in designing in-situ interventions and nudging users toward more effective search terms and paths [52, 87]. Apart from individual-level factors, researchers have also explored the effect of *social factors* on users' attitudes and behaviors (e.g. friends and colleagues know about the decision) [59]. Adding and changing social consequences has been demonstrated as an effective technique in promoting green lifestyle and health diet [30, 46], and may also be leveraged in encouraging critical thinking, active reflection on possible biased judgments on recommendations, and the acceptance of diverse opinions and perspectives. Due to the diverse nature of human biases, the specific intervention and nudging methods need to be customized and adjusted in real-time according to involved user characteristics and the in-situ estimated risks of individual biases. Also, the possible interactions and mutual-reinforcements between multiple biases need to be considered in designing a set of nudging techniques.

Based on the discussions above, Figure 3 summarizes the tasks associated with enhancing two-sided fairness. Similar to the three-stage structure of data and algorithmic debiasing, human-side fairness can also be achieved and enhanced at multiple stages. Specifically, systems could proactively estimate the potential risks of biases based on the knowledge about users learned from previous interaction data and the structure of current SERP and recommendation list. With the estimated risk levels, systems can decide the specific actions (e.g. re-ranking, adaptive intervention, digital nudging) to take in following interactions. In addition, when biased behaviors and judgments occur, systems can adjust the ranking algorithms and recommendation strategies accordingly, correct the biased results, and remove contextual triggers that could cause other biases for the user. Once the interaction is completed, the system may still provide post-interaction interventions (e.g. summarize and present possible biases extracted from whole-session interactions and implicit feedback) in order to at least partially address human bias in decision-making (e.g. making personal health decision, choosing recommended products to purchase, deciding which applicant to interview or hire).

4 PRACTICAL APPLICATIONS AND POTENTIAL CHALLENGES

When interacting with information systems and making decisions under uncertainty, users should be protected from the bias and discrimination that emerge from both system biases associated with sensitive attributes, but also the negative impact of human biases triggered by both individual traits and contextual factors. Our two-sided fairness framework could be applied in a broad range of search and recommendation interaction scenarios.

4.1 Practical Applications

The two-sided fairness concept offers a new perspective for evaluating users' interactions with systems of varying modalities. For instance, when evaluating the fairness of *Conversational Search and Recommendation systems*, apart from the observable biases in system responses, researchers should also investigate the signals of cognitive biases in utterances and interactions. In addition, systems can predict and clarify potential biases (e.g. misleading beliefs and unrealistic expectations regarding certain products) with users by promoting recommended questions, asking clarifying questions, and analyzing users' reactions. Once a potential bias is identified, the system could push certain reminders or alerts to the user before an unfair decision is made. Similarly, in traditional recommendation scenarios, systems can proactively analyze the item list to be recommended and estimate the risk of cognitive biases being triggered by an individual item (e.g. reference dependence, confirmation and status quo bias) or combination of biases (e.g. loss and risk aversion, decoy effect). With the information about both the user and built-in recommendation algorithms, systems can develop and adaptively adjust the models that predict users' vulnerability to different types of biases and provided personalized solutions. For instance, a system could apply a reinforcement learning (RL) based approach that can offer iteratively optimized ranking and recommendations based on identified *bias states*, which could be represented as varying estimated risks of biases at the moment.

Similarly, the two-sided fairness approach could also be employed in measuring and mitigating algorithmic and human biases in social media platforms. For instance, systems should include the affordance and components that allow them to proactively identify the triggered cognitive biases that may increase a user's chance of receiving and accepting certain health misinformation. Predicting and identifying biased behavior and judgments may need to involve two models: 1) a *global model* that captures and dismantles the structure of recommendations triggering biases in information evaluation (e.g. identifying and removing potential decoy results that could trigger the acceptance of low-quality or irrelevant information); 2) a *personalized model* that covers individual characteristics for assessing the risk of human bias (e.g. extracting in-situ reference points and expectations from past similar experiences

and on-going interactions and estimate perceived gains and costs from interactions).

Apart from system-initiated intervention and intelligent nudging, addressing two-sided biases, particularly human biases, could also be achieved through enhancing users' *algorithmic literacy and awareness* [8, 44, 65]. For instance, librarians as the traditional information gatekeepers can design and implement education programs and tools for community members who are not familiar with search and recommendation algorithms or are vulnerable to the bias and discrimination emerging from both system and human biases. Library and Information Science (LIS) professionals may provide proactive support for users engaging in complex, black-box recommender systems and help them understand the decisions of algorithms and the impacts of their own real-time behaviors and feedback on the scope and focus of recommendations. In addition, incorporating the two-sided fairness framework and practices into *Information Search Education* could improve users' awareness of their potential biases when interacting with the results from search and recommender systems and facilitate effective, fair decision-making.

4.2 Potential Challenges

While the two-sided fairness framework can extend the scope of promoting fairness in IR, RS and beyond, applying the approach also involves additional challenges that need to be studied and addressed. Many of the potential challenges are associated with the preparation tasks to be completed for estimating human biases (see Section 3.5).

Specifically, regarding T1, the prediction results will serve as the ground truth label for measuring the fairness of equal probability in achieving desired or optimal outcomes. This is challenging mainly for three reasons: 1) predicting desired outcomes requires the knowledge about user intentions and the nature of motivating task, both of which are difficult to obtain in real-time interactions according to existing relevant research [49]; 2) In some scenarios where the motivating tasks are complex and ill-defined, users themselves may not have a clear goal or desired outcome, leading to difficulty for evaluating fairness from human side; 3) In contrast to the pre-defined protected attributes and fairness goals in algorithmic side, users' intentions and desired outcomes may change over time, which calls for an adaptive, context-dependent approach to assessing fairness. Apart from the technical difficulties, predicting desired outcomes and estimating the risks of biases (i.e. T2) would require the information about users' background and prior experiences under similar problems, which may lead to ethical issues and privacy concerns. To address this challenge, system designers have to 1) protect and restrict the usage of interaction history data and human bias data in model training and fairness evaluation, and 2) enable users to be aware of and have control over the collection, usage, and processing of the data regarding their potential cognitive and perceptual biases. In addition, certain restrictions and regulations should be implemented for better managing the data reuse and replication experiments in human-centered fairness evaluation.

With respect to T3, learning accurate, useful bias-aware user models may require data regarding user behaviors both within and outside interactive information systems (e.g. users' offline purchase decisions in supermarkets under the effect of changing reference prices and decoy options; users' existing understanding and preferences on a foreign policy topic after reading a political science textbook). In addition to the challenges in data collection, users themselves may have difficulty in labeling their own biased behaviors in naturalistic settings as most of the cognitive biases operate unconsciously in information evaluation and decision-making scenarios, causing obstacles for training and testing bias-aware user models. Also, since the fairness of decisions may also be affected by the factors outside search and recommendations (e.g. users' existing biases and beliefs, available support from domain experts, time constraints), changing ranking algorithms, recommendation mechanisms, and interface presentations only may not guarantee a successful transition of information fairness to the tangible fairness in task performances and decision-making.

5 NEW QUESTIONS AND DIRECTIONS

Under the two-sided fairness framework, we propose a series of new questions and directions for encouraging discussions on related problems and inspiring future research on user-centered fairness evaluation.

5.1 Understanding Biases from Different Sources

As the basis for measuring and promoting fairness, researchers need to investigate biases from varying sources and characterize the implicit interactions among them. We can start with addressing following research questions (RQs):

- **RQ1**: How are different user biases triggered by their previous experiences, system outputs, and individual characteristics in users' interactions with search and recommender systems?
- **RQ2**: How do user biases interact with system biases at different phases of interactions, such as query reformulation, browsing and examination of search results and recommendations, clicking and evaluation?
- **RQ3**: To what extent does the distribution of user biases vary across different tasks and systems?

Addressing the first three RQs would require researchers to conduct extensive user studies and carefully examine the connections of individual biases to users, tasks, and systems. Particularly, it is critical to differentiate individual biases from in-situ natural preferences and enhance users' awarenss of potential risks without intervening their tasks. This could be challenging especially when different types of biases are correlated with each other, generating a mixed effect on judgments of information items and post-interaction decision-making. Knowledge learned under these RQs will offer an empirical basis for estimating the risks of biases in real-time interactions.

5.2 Evaluating Two-Sided Fairness

On human-sided fairness evaluation, our work defines the fairness goal without specifying individual fairness measures. Determining the specific measures would require answers to at least two RQs:

- **RQ4**: How do different user biases lead to biased judgments and unfair decisions?
- **RQ5**: How can we evaluate system fairness in addressing the negative impact of human biases under different desired outcomes and intentions of varying types?

Differing from algorithmic debiasing where the goal can be quantified beforehand once the protected attributes are defined, different human biases may be coupled with varying types of behaviors and desired outcomes, which require customized fairness measures and constraints on system training and evaluation. Therefore, RQ4 and RQ5 highlight the connections of human fairness measures to user behaviors, decisions and goals of interactions, aiming to clarify the role of human biases in interaction processes. Findings under the two RQs may result in a bias-aware user model that characterizes user search behavior and decision-making patterns under varying biases, and separate fairness metrics for evaluation under varying intentions and desired outcomes.

5.3 Enhancing Two-Sided Fairness

Moving towards enhancing two-sided fairness, researchers need to examine the usefulness and appropriateness of the available tools at hand. Aligned with the discussions presented in Section 4.1, we propose following two RQs as a starting point for this direction of research:

- **RQ6**: How can we enhance two-sided fairness and address varying types of biases using re-ranking, intervention, and intelligent nudging techniques?
- **RQ7**: How can we we enhance two-sided fairness and address varying types of biases through improving users' algorithmic literacy and awareness of human and system biases?

Differing from previously asked causal-inference questions (e.g. RQ1, RQ4), RQ6 and RQ7 are closely related to application-oriented practical questions and may yield highly contextual-dependent answers in field studies. For instance, the effectiveness of specific interventions, nudging techniques, and algorithmic literacy education programs may vary significantly across different types of systems and populations from varying background. Therefore, studying these two RQs may also involve fairness issue - the unique traits, needs, and challenges of different community members and groups should be fully considered and equally represented when testing intervention tools and education programs. Also, in practical application, researchers and system designers should balance the autonomy of users and the role of recommendations, and evaluate a broad range of approaches to enhancing two-sided fairness with users (e.g. from in-situ reminders of possible biased judgments and suggestions of search tactics to proactive re-ranking and recommendations based on estimated risks).

5.4 Ethical Challenges and Data Reusability

Similar to other user-centered evaluation studies, research on two-sided fairness will involve the sensitive, expensive, and time-consuming process of collecting labels and signals regarding user features (in this case, human-side biases and fairness). As a result, researchers need to face a set of ethical and practical challenges. The exploration on this problem space may start with two RQs:

- **RQ8**: How can we measure and promote two-sided fairness and also protect users' private information regarding individual biases and previous experiences?
- **RQ9**: How can we effectively reuse the data regarding two-sided biases and fairness and amortizing the true cost of user experiments?

RQ8 could be address by adding additional privacy protection constraints on model training and system outputs, and design punishments (e.g. significantly reducing the evaluation score) when the risk of bias data leakage is captured. Regarding RQ9, researchers need to develop a standard framework for guiding data curation and sharing and assessing the reusability of behavior and annotation datasets collected from individual user studies [26, 50]. Effective data reuse would allow researchers to develop and meta-evaluate the effectiveness of two-sided fairness measures across varying dtasets, systems, and populations.

5.5 From User to People Interacting with Information

Apart from the specific RQs and new directions presented above, our long-term vision is to studying users as people interacting with information, rather than as agents operating in systems disconnected from specific tasks and socio-technical contexts. The main idea behind this vision is that we cannot expect people to leave their specific contexts for interacting with search and recommender systems and act as "users" in the way we assumed [9]. Instead, people's interactions with information systems should be characterized and evaluated in *contexts*.

Aligned with this idea, our two-sided fairness framework goes beyond traditional system fairness measures that sets clear boundaries between algorithms and users, and investigate the concept of human-centered fairness that reconnect users and their biases with factors from their contexts and problematic situations. In this sense, studying and implementing the two-sided fairness evaluation will not only expand the scope of research on fairness in IR, RS and Human-AI interaction in general, but also contribute to the general efforts on bringing users interacting with information and systems back to their contexts in human-centered computing research.

6 CONCLUSION

As artificial intelligence (AI) assisted search and recommender systems have become ubiquitous in workplaces and everyday lives, understanding and accounting for *fairness* has gained increasing attention in the design and evaluation of such systems. While there is a growing body of computing research on measuring system fairness and biases associated with data and algorithms, the impact of human biases that go beyond traditional machine learning (ML) pipelines still remain understudied. To address this challenge, our study extends the concept of fairness to cover the effects and measurements of both human bias and system bias embedded in data and algorithms, as well as the possible interactions between them. Also, we propose new two-sided evaluation goals and methods that can examine the performance of search and recommender systems in addressing and proactively reducing the impacts of both human and system biases. In addition, our paper synthesizes relevant re-ranking, intervention and nudging techniques that could

potentially mitigate the risks of one or both types of biases, and identifies technical and ethical challenges as well as new directions for future fairness-oriented evaluation research in IR and RS.

We hope that the new insights, perspectives and questions we presented on the two-sided fairness problem can incite fruitful discussions in CHIIR community and also encourage information researchers and scientists to further push the boundaries of fairness evaluation research from a human-centered perspective.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (NSF) award IIS-2106152.

REFERENCES

[1] Deena Abul-Fottouh, Melodie Yunju Song, and Anatoliy Gruzd. 2020. Examining algorithmic biases in YouTube's recommendations of vaccine videos. *International Journal of Medical Informatics* 140 (2020), 104175.

[2] 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.

[3] Qingyao Ai, Tao Yang, Huazheng Wang, and Jiaxin Mao. 2021. Unbiased learning to rank: Online or offline? *ACM Transactions on Information Systems (TOIS)* 39, 2 (2021), 1–29.

[4] Elliot Aronson. 1969. The theory of cognitive dissonance: A current perspective. In *Advances in experimental social psychology*. Vol. 4. Elsevier, 1–34.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[6] Leif Azzopardi. 2021. Cognitive biases in search: a review and reflection of cognitive biases in Information Retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval*. 27–37.

[7] Aakriti Bajracharya, Utsab Khakurel, Barron Harvey, and Danda B Rawat. 2023. Recent Advances in Algorithmic Biases and Fairness in Financial Services: A Survey. In *Proceedings of the Future Technologies Conference*. Springer, 809–822.

[8] Abigail Bakke. 2020. Everyday Googling: Results of an observational study and applications for teaching algorithmic literacy. *Computers and Composition* 57 (2020), 102577.

[9] Nicholas J Belkin. 2016. People, interacting with information. In *ACM SIGIR Forum*, Vol. 49. ACM New York, NY, USA, 13–27.

[10] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[11] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409* (2017).

[12] Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in BERT. *Cognitive Computation* 13, 4 (2021), 1008–1018.

[13] 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 & development in information retrieval*. 405–414.

[14] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. PMLR, 149–159.

[15] Tyler Brown and Jiqun Liu. 2022. A reference dependence approach to enhancing early prediction of session behavior and satisfaction. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*. 1–5.

[16] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–15.

[17] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*. 621–630.

[18] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*. 339–348.

[19] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89.

[20] Andrew Collins, Dominika Tkaczyk, Akiko Aizawa, and Joeran Beel. 2018. Position bias in recommender systems for digital libraries. In *International Conference on Information*. Springer, 335–344.

[21] Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big data* 5, 2 (2017), 120–134.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[23] 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.

[24] Cynthia Dwork and Christina Ilvento. 2018. Group fairness under composition. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency (FAT* 2018)*.

[25] Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 162–170.

[26] Maria Gäde, Marijn Koolen, Mark Hall, Toine Bogers, and Vivien Petras. 2021. A Manifesto on Resource Re-Use in Interactive Information Retrieval. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 141–149.

[27] Ruoyuan Gao, Yingqiang Ge, and Chirag Shah. 2022. FAIR: Fairness-aware information retrieval evaluation. *Journal of the Association for Information Science and Technology* (2022).

[28] Yingqiang Ge, Shuya Zhao, Honglu Zhou, Changhua Pei, Fei Sun, Wenwu Ou, and Yongfeng Zhang. 2020. Understanding echo chambers in e-commerce recommender systems. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 2261–2270.

[29] Itzhak Gilboa and David Schmeidler. 1995. Case-based decision theory. *The quarterly Journal of economics* 110, 3 (1995), 605–639.

[30] Diogo Gonçalves, Pedro Coelho, Luis F Martinez, and Paulo Monteiro. 2021. Nudging consumers toward healthier food choices: A field study on the effect of social norms. *Sustainability* 13, 4 (2021), 1660.

[31] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, Vol. 1. Barcelona, Spain, 2.

[32] Huifeng Guo, Jinkai Yu, Qing Liu, Ruiming Tang, and Yuzhou Zhang. 2019. PAL: a position-bias aware learning framework for CTR prediction in live recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 452–456.

[33] Calin Gurau. 2015. The effect of marketing promotions on customers' cognitive biases. In *The Proceedings of the International Conference" Marketing-from Information to Decision"*. Babes Bolyai University, 48.

[34] Kevin Hamilton, Karrie Karahalios, Christian Sandvig, and Motahhare Eslami. 2014. A path to understanding the effects of algorithm awareness. In *CHI'14 extended abstracts on human factors in computing systems*. 631–642.

[35] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[36] Chia-Fen Hsu, Lee Propp, Larissa Panetta, Shane Martin, Stella Dentakos, Maggie E Toplak, and John D Eastwood. 2018. Mental effort and discomfort: Testing the peak-end effect during a cognitively demanding task. *PloS one* 13, 2 (2018), e0191479.

[37] Ben Hutchinson and Margaret Mitchell. 2019. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 49–58.

[38] Rolf Jagerman, Harrie Oosterhuis, and Maarten de Rijke. 2019. To model or to intervene: A comparison of counterfactual and online learning to rank from user interactions. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*. 15–24.

[39] Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *American economic review* 93, 5 (2003), 1449–1475.

[40] Daniel Kahneman, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

[41] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*. PMLR, 2564–2572.

[42] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2019. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 100–109.

[43] Miles S Kimball. 1993. Standard risk aversion. *Econometrica: Journal of the Econometric Society* (1993), 589–611.

[44] Abby Koenig. 2020. The algorithms know me and i know them: using student journals to uncover algorithmic literacy awareness. *Computers and Composition* 58 (2020), 102611.

[45] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[46] Loni Ledderer, Marianne Kjær, Emilie Kirstine Madsen, Jacob Busch, and Antoinette Fage-Butler. 2020. Nudging in public health lifestyle interventions: a systematic literature review and metasynthesis. *Health Education & Behavior* 47, 5 (2020), 749–764.

[47] Kwan Min Lee, Younbo Jung, and Clifford Nass. 2011. Can user choice alter experimental findings in human–computer interaction?: Similarity attraction versus cognitive dissonance in social responses to synthetic speech. *Intl. Journal of Human–Computer Interaction* 27, 4 (2011), 307–322.

[48] Michelle Seng Ah Lee and Luciano Floridi. 2021. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines* 31, 1 (2021), 165–191.

[49] Jiqun Liu. 2021. Deconstructing search tasks in interactive information retrieval: A systematic review of task dimensions and predictors. *Information Processing & Management* 58, 3 (2021), 102522.

[50] Jiqun Liu. 2022. Toward Cranfield-inspired reusability assessment in interactive information retrieval evaluation. *Information Processing & Management* 59, 5 (2022), 103007.

[51] Jiqun Liu and Fangyuan Han. 2020. Investigating reference dependence effects on user search interaction and satisfaction: A behavioral economics perspective. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1141–1150.

[52] Jiqun Liu, Yiwei Wang, Soumik Mandal, and Chirag Shah. 2019. Exploring the immediate and short-term effects of peer advice and cognitive authority on Web search behavior. *Information Processing & Management* 56, 3 (2019), 1010–1025.

[53] Ramona Ludolph and Peter J Schulz. 2018. Debiasing health-related judgments and decision making: a systematic review. *Medical Decision Making* 38, 1 (2018), 3–13.

[54] David J Malenka, John A Baron, Sarah Johansen, Jon W Wahrenberger, and Jonathan M Ross. 1993. The framing effect of relative and absolute risk. *Journal of general internal medicine* 8, 10 (1993), 543–548.

[55] Jiaxin Mao, Yiqun Liu, Noriko Kando, Cheng Luo, Min Zhang, and Shaoping Ma. 2018. Investigating result usefulness in mobile search. In *European Conference on Information Retrieval*. Springer, 223–236.

[56] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[57] Alessandro B Melchiorre, Navid Rekabsaz, Emilia Parada-Cabaleiro, Stefan Brandl, Oleg Lesota, and Markus Schedl. 2021. Investigating gender fairness of recommendation algorithms in the music domain. *Information Processing & Management* 58, 5 (2021), 102666.

[58] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems (TOIS)* 27, 1 (2008), 1–27.

[59] Robert Münscher, Max Vetter, and Thomas Scheuerle. 2016. A review and taxonomy of choice architecture techniques. *Journal of Behavioral Decision Making* 29, 5 (2016), 511–524.

[60] Alamir Novin and Eric Meyers. 2017. Making sense of conflicting science information: Exploring bias in the search engine result page. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*. 175–184.

[61] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[62] Lihong Peng, Yi Guo, and Dehua Hu. 2021. Information framing effect on public's intention to receive the COVID-19 vaccination in China. *Vaccines* 9, 9 (2021), 995.

[63] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.

[64] Donald A Redelmeier, Joel Katz, and Daniel Kahneman. 2003. Memories of colonoscopy: a randomized trial. *Pain* 104, 1-2 (2003), 187–194.

[65] Michael Ridley and Danica Pawlick-Potts. 2021. Algorithmic literacy and the role for libraries. *Information technology and libraries* 40, 2 (2021).

[66] William Samuelson and Richard Zeckhauser. 1988. Status quo bias in decision making. *Journal of risk and uncertainty* 1, 1 (1988), 7–59.

[67] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[68] Christoph Schneider, Markus Weinmann, and Jan Vom Brocke. 2018. Digital nudging: guiding online user choices through interface design. *Commun. ACM* 61, 7 (2018), 67–73.

[69] Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S Lee, and William Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 623–632.

[70] Christina Schwind and Jürgen Buder. 2012. Reducing confirmation bias and evaluation bias: When are preference-inconsistent recommendations effective–and when not? *Computers in Human Behavior* 28, 6 (2012), 2280–2290.

[71] Intan Sherlin, Ferry Siswadhi, and Elex Sarmigi. 2020. Analysing the Decoy Effect on Online Product Purchasing Preference: an experimental study. In *6th Annual International Conference on Management Research (AICMaR 2019)*. Atlantis Press, 125–130.

[72] Herbert A Simon. 1955. A behavioral model of rational choice. *The quarterly journal of economics* 69, 1 (1955), 99–118.

[73] Cass R Sunstein. 2016. The council of psychological advisers. *Annual Review of Psychology* 67 (2016), 713–737.

[74] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* 2 (2019), 8.

[75] Richard H Thaler. 2016. Behavioral economics: Past, present, and future. *American economic review* 106, 7 (2016), 1577–1600.

[76] Georgios Theocharous, Jennifer Healey, Sridhar Mahadevan, and Michele Saad. 2019. Personalizing with human cognitive biases. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. 13–17.

[77] Samuel P Trethewey. 2019. Medical misinformation on social media: cognitive bias, pseudo-peer review, and the good intentions hypothesis. *Circulation* 140, 14 (2019), 1131–1133.

[78] Jennifer S Trueblood and Jonathan C Pettibone. 2017. The phantom decoy effect in perceptual decision making. *Journal of Behavioral Decision Making* 30, 2 (2017), 157–167.

[79] Amos Tversky and Daniel Kahneman. 1991. Loss aversion in riskless choice: A reference-dependent model. *The quarterly journal of economics* 106, 4 (1991), 1039–1061.

[80] Elmira van den Broek, Anastasia Sergeeva, and Marleen Huysman. 2019. Hiring algorithms: An ethnography of fairness in practice. (2019).

[81] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*. IEEE, 1–7.

[82] Chao Wang, Yiqun Liu, Min Zhang, Shaoping Ma, Meihong Zheng, Jing Qian, and Kuo Zhang. 2013. Incorporating vertical results into search click models. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. 503–512.

[83] 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.

[84] Yifan Wang, Weizhi Ma, Min Zhang*, Yiqun Liu, and Shaoping Ma. 2022. A survey on the fairness of recommender systems. *ACM Journal of the ACM (JACM)* (2022).

[85] Roberto A Weber and Colin F Camerer. 2006. " Behavioral experiments" in economics. *Experimental Economics* 9, 3 (2006), 187.

[86] Chunhua Wu and Koray Cosguner. 2020. Profiting from the decoy effect: A case study of an online diamond retailer. *Marketing Science* 39, 5 (2020), 974–995.

[87] Yusuke Yamamoto and Takehiro Yamamoto. 2018. Query priming for promoting critical thinking in web search. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 12–21.

[88] Tao Zhang and David Zhang. 2007. Agent-based simulation of consumer purchase decision-making and the decoy effect. *Journal of business research* 60, 8 (2007), 912–922.